UNIVERSITY OF CANTERBURY

DOCTORAL THESIS

Development and Evaluation of a Fully Expressive Avatar Control System for Communication and Collaboration in Virtual Reality

Author:
Yuanjie WU

Supervisor:
Prof. Robert W. LINDEMAN
Dr. Sungchul JUNG
Dr. Simon HOERMANN

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Human Interface Technology Lab New Zealand

September 29, 2020

Declaration of Authorship

I, Yuanjie WU, declare that this thesis titled, "Development and Evaluation of a Fully Expressive Avatar Control System for Communication and Collaboration in Virtual Reality
" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Yuanjie Wu

Date: September 29, 2020

University of Canterbury

Abstract

Human Interface Technology Lab New Zealand

Doctor of Philosophy

Development and Evaluation of a Fully Expressive Avatar Control System for Communication and Collaboration in Virtual Reality

by Yuanjie WU

In current virtual reality (VR) systems, making avatars expressive in non-verbal behavior (body movement, hand gestures, facial expressions, and eye gaze) is difficult due to complex sensory integration and data fusion requirements. Hence, the social impacts of avatar expressiveness in terms of non-verbal behavior have not yet been thoroughly investigated in shared virtual environments.

I presented a novel expressive avatar system and conducted user studies to evaluate the system. The goal of the expressive avatar system was to develop it using off-the-shelf technology that provides an accurate, contactless, and natural interaction experience. The system consists of three parts. A customized multiple depth-camera system (MS Kinect v2) was developed for full-body tracking regardless of the user's orientation. A multiple short-range depth-camera (Leap Motion) system was developed to provide a controller-free experience for enlarged natural hand gesture interaction. An avatar integration system was developed for avatar control and rendering with customized facial expressions such as the mouth and eye movement. To integrate these three systems I have designed a highly expressive avatar control system framework that includes a novel adaptive weighting data fusion method, an enlarged usable hands fusion method, and a robust avatar control algorithm.

The system was evaluated in two user studies. The first user study explored the effects of a depth-sensor-based avatar system on social behavior and performance in the single-user simulated communication scenario. The second user study explored the impact of different levels of avatar expressiveness on collaboration behavior through a virtual charade game. My results demonstrated that a highly expressive avatar control system increased virtual body ownership and agency, improved user experience, and produced better non-verbal performance in single-user simulated communication scenarios. Furthermore, users felt a deeper sense of social presence and more attraction when interacting with a user who was using a highly expressive avatar in collaborative tasks.

In summary, this thesis provides novel technical contributions on how to develop highly expressive avatar systems for communication and collaboration in shared virtual environments and evidence of how highly expressive avatars can benefit users in these scenarios.

Acknowledgements

During my PhD, I got lots of help from the people around me. At this stage, I would like to show my deepest gratitude to them.

I wish to express my sincere appreciation to my academic supervisor, Professor Rob Lindeman, thanks for his support, help, and patience. I appreciate him taking charge of me as his PhD student when I changed my research direction, and providing the project funding support when I had troubles in my research and life. It would be impossible for me to finish my PhD research without his persistent help. I would like to thank my co-supervisor, Dr. Sungchul Jung, for the help with my research. He gave me many suggestions and feedback on my user study. He always quickly responded when I was asking him for help with my research. Thanks to my co-supervisor, Dr. Simon Hoermann. He gave me many suggestions and feedback on the thesis. Thanks to the academic visitor Yu Wang, with enlightening suggestions and support, I make the avatar system more robust. Also, thanks to the academic visitor Associate Professor Christoph Borst. He gave me lots of help in mathematics for the avatar system. Thanks to my friends Huidong Bai, Hao Chen, Lei Gao, Kris Tong, and other colleagues in the HIT Lab for sharing the research experience and organizing the activities in the leisure time. I also would like to thank all the examiners for spending their valuable time to read and evaluate my thesis.

I wish to acknowledge the support and great love of my parents for their encouragement throughout my study. Thank you for always believing me and supporting me when I make crucial decisions. I want to thank my mother and father in law for understanding and encouraging in the last six years. Finally, huge thanks to my wife, Lingzi Xue. Thanks for your accompanying and encouragement for so many years.

Deputy Vice-Chancellor's Office
Postgraduate Research Office

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Chapter 3 *of the thesis is a reproduction of the work undertaken and published in collaboration with Lei Gao, Simon Hoermann, and Robert W. Lindeman. Results from this work have been presented at the following conference.*

10th International Conference on Virtual Worlds and Games for Serious Applications (VS Games 2018)

Yuanjie Wu, Lei Gao, Simon Hoermann, and Robert W. Lindeman. (2018, September). Towards Robust 3D Skeleton Tracking Using Data Fusion from Multiple Depth Sensors. In 2018 10th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games) (pp. 1-4). IEEE.

Please detail the nature and extent (%) of contribution by the candidate:

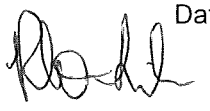
Rob Lindeman and Simon Hoermann helped me on the study's methodology and gave me good feedback on the user study design. Lei Gao helped me design part of the data fusion system. Rob Lindeman helped review and edit the paper and gave good feedback after the paper draft was finished. My contribution is more than 90%

Certification by Co-authors:

If there is more than one co-author then a single co-author can sign on behalf of all

The undersigned certifies that:

- The above statement correctly reflects the nature and extent of the Doctoral candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name:	Signature:	Date:
ROBERT LINDEMAN		2020-09-25

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Deputy Vice-Chancellor's Office
Postgraduate Research Office

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Chapter 4 *of the thesis is a reproduction of the work undertaken and published in collaboration with Yu Wang, Sungchul Jung, Simon Hoermann, Robert W. Lindeman. Results from this work have been presented at the following journal.*

June 2019 Entertainment Computing, Elsevier

Yuanjie Wu, Yu Wang, Sungchul Jung, Simon Hoermann, Robert W. Lindeman. (2019). Towards an articulated avatar in VR: Improving body and hand tracking using only depth cameras. Entertainment Computing, 31, 100303.

Please detail the nature and extent (%) of contribution by the candidate:

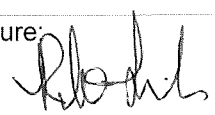
Rob Lindeman, Sungchul Jung, and Simon Hoermann helped me on the study's methodology and gave me good feedback on the user study design. Yu Wang helped me implement the experiment and analyse part of the data. Rob Lindeman helped review and edit the paper and gave good feedback after the paper draft was finished. My contribution is more than 85%

Certification by Co-authors:

If there is more than one co-author then a single co-author can sign on behalf of all

The undersigned certifies that:

- The above statement correctly reflects the nature and extent of the Doctoral candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name:	Signature:	Date:
ROBERT LINDEMAN		2020-09-25

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Deputy Vice-Chancellor's Office
Postgraduate Research Office

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Please indicate the chapter/section/pages of this thesis that are extracted from co-authored work and provide details of the publication or submission from the extract comes:

Chapter 5 *of the thesis is a reproduction of the work undertaken and published in collaboration with Yu Wang, Sungchul Jung, Simon Hoermann, Robert W. Lindeman. Results from this work have been presented at the following conference.*

25th ACM Symposium on Virtual Reality Software and Technology(VRST)

Yuanjie Wu, Yu Wang, Sungchul Jung, Simon Hoermann, Robert W. Lindeman. (2019, November). Exploring the Use of a Robust Depth-sensor-based Avatar Control System and its Effects on Communication Behaviors. In 25th ACM Symposium on Virtual Reality Software and Technology (pp. 1-9).

Please detail the nature and extent (%) of contribution by the candidate:

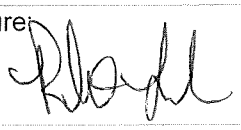
Rob Lindeman, Sungchul Jung, and Simon Hoermann helped me on the study's methodology and gave me good feedback on the user study design. Yu Wang helped me implement the experiment. Rob Lindeman and Sungchul Jung helped review and edit the paper and gave good feedback after the paper draft was finished. My contribution is more than 90%

Certification by Co-authors:

If there is more than one co-author then a single co-author can sign on behalf of all

The undersigned certifies that:

- The above statement correctly reflects the nature and extent of the Doctoral candidate's contribution to this co-authored work
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text

Name:	Signature:	Date:
ROBERT LINDEMAN		2020-09-25

Contents

List of Figures

List of Tables

List of Abbreviations

VR	**V**irtual **R**eality
VE	**V**irtual **E**vironment
SVE	**S**hared **V**irtual **E**nvironment
VC	**V**erbal **C**ommunication
NVC	**N**on-**V**erbal **C**ommunication
LMC	**L**eap **M**otion **C**ontroller
ICP	**I**terative **C**losest **P**oint
LSF	**L**east **S**qures **F**itting
IMU	**I**nertial **M**easurement **U**nits
JDL	**J**oint **D**irectors of **L**aboratories (JDL)
LRS	**L**eft **R**ight **S**wap
HMD	**H**ead **M**ounted **D**isplay
IK	**I**nverse **K**inematics
SWC	**S**ub-region **W**eight **C**alculation
AWC	**A**daptive **W**eight **C**alculation
IR	**I**nfrared **R**adiation
OSC	**O**pen **S**ound **C**ontrol
CB-ACS	**C**ontroller **B**ased **A**vatar **C**ontrol **S**ystem
DSB-ACS	**D**epth **S**ensor **B**ased **A**vatar **C**ontrol **S**ystem
IPQ	**I**group **P**resence **Q**uestionnaire
FPP	**F**irst **P**erson **V**iew
TPP	**T**hird **P**erson **V**iew
RWV	**R**eal **W**orld **V**iew
VBOI	**V**irtual **B**ody **O**wnership **I**llusion
GP	**G**eneral **P**resence
SP	**S**patial **P**resence
FOV	**F**ield **O**f **V**iew
PCL	**P**oint **C**loud **L**ibrary
SVD	**S**ingular **V**alue **D**ecomposition
UI	**U**ser **I**nterface
HEA	**H**ighly **E**xpressive **A**vatar
LEA	**L**ow **E**xpressive **A**vatar
UDP	**U**ser **D**atagram **P**rotocol

Chapter 1

Introduction

Virtual Reality (VR) is a computer technology that can provide people an experience of a virtual world generated by a computer [121]. The Virtual Environment (VE) can be effectively experienced and interacted with as if it were real and responsive [58]. With the development of applicable technologies of computer vision, graphic computing, and wide-range tracking sensors, researchers and developers can create a variety of application scenarios for education, training, social interaction, and games [111]. Reconstructing and replicating the real-world scene, objects, and detailed interaction through sensory data, VR is capable of making users spend a long time immersing themselves in the virtual environment. Researchers and developers strive to provide a realistic experience for people with user-centred design. Current VR applications not only focus on interacting with objects or the environment but also stress communicating and collaborating with other users in the shared virtual environment (SVE) with different geographic locations. As a medium, virtual representation is playing an essential role in connecting the user and the virtual world. An avatar is a virtual character that represents the user in the virtual world. The avatar's action can be controlled by a human in real-time [90]. Existing technologies enable computer-generated entities to mimic both the appearance and behaviors of humans [18, 30], and the user can interact with the virtual world through the eyes of the virtual character from the first person point of view. Avatar realism is often used to measure the quality of the avatar, which can be divided into appearance and behavior realism [119]. With the help of modeling software, the user can

get the cartoonish character or even a photo-realistic version using the 3D reconstruction and capture technology. Most previous research focuses on the effect of appearance or form realism that avatar can bring in the single user or multiple user scenario. Due to the sensory technology, the avatar control system (integration of non-verbal behavior) is limited in the current immersive systems. For example, the user needs to wear a cumbersome and expensive tracking suit for accurate full-body tracking. Additionally, they have to use tracking gloves or controllers for hand gesture-based interaction. Few systems can provide a natural and integral interaction solution in VR. Although we still have demand to improve the appearance of realism [15], the impacts of behavioral expressiveness have not yet been systematically investigated in the virtual environment with fully embodied avatars for communication and collaboration.

1.1 Motivation

Verbal and non-verbal behavior are two of the main components of communication and are essential tools for mutual understanding [79]. Verbal communication (VC) is a direct way to express thoughts and ideas to other humans by using words. In contrast, non-verbal communication (NVC), such as body language, could help reduce the risk of misunderstandings. NVC can take many forms [45], such as body movements, body posture, hand gestures, eye contact (eye gaze direction, eye blinking), tone, voice pitch, and facial expressions. An embodied avatar is the medium that enables people to interact and employ both VC and NVC in VE. With modern tracking technology, the player generally implements NVC through a virtual character whose behavior is captured by peripheral devices. The user can view the virtual world through the avatar's eyes, and the avatar movement reflects their body movement [115]. The realism of avatars in terms of form and behavior is important for communication and collaboration in VEs [127, 42]. Most previous work has been done on visual fidelity [71, 126], and avatar appearance does influence interactions in all shared VEs [88, 108]. Several researchers

shared the alternative viewpoint that behavioral fidelity is a higher priority. Salinäs and Eva-Lotta [105] argued that realistic appearance is secondary to support of body posture, gesture, and object manipulation in collaboration tasks. Blascovich [13] and Swinth [124] also argued that photographic realism is less important than behavioral realism. This leads to the question that motivates this research:*"How and to what extent does behavioral realism affect the user in terms of presence, agency, body ownership illusion, performance, and social presence in single-user or multi-user scenarios?"*.

The quality of presented behavioral realism depends on the avatar system, but early avatar control systems could not provide the complete embodied experience. Tracking technology limitations such as tracking accuracy, tracking range, data fusion, avatar control algorithm, and avatar rendering can cause missed information channels such as body movement, hand gesture, and facial expressions. The current technology is not able to capture all the non-verbal behavior from just a single peripheral input device. In order to address the problem, a combination of heterogeneous tracking systems has been suggested [99].

The body movement of the user is the primary source of data for the avatar. To get a high quality embodied experience, a motion capture suit is widely used in the avatar related research [65, 101, 117], which is expensive and cumbersome. In contrast, consumer VR devices such as the *Oculus Rift*[1] or *HTC Vive*[2] with spatial controllers are alternative solutions for tracking parts of the body. Most VR applications are based on this three-tracking-point (one HMD plus two controllers) solution, which only supports "floating" avatars, such as *Facebook spaces*[3], *VR Chat*[4], and *Mozilla Hubs*[5]. Extra trackers are required along with specific inverse kinematic software if the player needs a full-body experience [21]. Compared to the HMD and controller solution, RGB-D camera-based body tracking is a contactless way that can

[1]https://www.oculus.com/rift/
[2]https://www.vive.com/nz/product/vive/
[3]https://www.facebook.com/spaces
[4]https://vrchat.com/
[5]https://hubs.mozilla.com/

provide more joint information. The combination of an RGB-D sensor and VR device is another solution to support body tracking without wearing tracking sensors [69]. Users can experience improved articulation control of their avatar using these approaches.

This thesis will present a highly expressive avatar control system with natural body movement and hand gesture tracking along with eye and mouth movement. Studies were designed and implemented to investigate the effect of this highly expressive avatar control system on user communication and collaboration behavior subjectively, and objectively.

1.2 Research Questions

The research questions listed in the following part are based on the research in the previous section.

- Q1: Avatar related research has been studied for decades. What can I learn from the previous research?
  - 1: What can be used to measure the effect of non-verbal behavior that is presented from the avatar system on communication and collaboration? What are the measurements that are used in the previous research?
  - 2: What technique or solutions could be used to present avatar tracking? What are the pros and cons of the approaches? Where is the room for improvement in terms of full body and hand movement tracking?
- Q2: While ensuring accuracy, how can a full-body tracking system be built without attaching tracking devices to the user that allows people to freely control their avatars?
- Q3: To interact in the virtual environment, how to provide a controller-free experience in VR to the player in terms of full-body tracking with natural hand gesture interaction?

- Q4: Compared to a controller-based avatar control system, how and to what extent can the integration of multiple depth sensors of the avatar control system can support communication?
- Q5: To provide smooth hands control in VR, how to enlarge the usable hand tracking area only using depth sensors?
- Q6: How and to what extent does a high level of non-verbal expressiveness support collaboration in a SVE?

1.3 Findings and Contributions

The work presented in this thesis makes a contribution to building highly expressive avatar control systems and explores the effects of multiple sensor-based avatar control systems on supporting communication and collaboration in VR. In particular, three important contributions and findings from the thesis are:

(a) A highly expressive avatar control system was built, which can support verbal and non-verbal behavior for communication and collaboration in single or multi-user scenarios.

- A novel framework and algorithm was proposed to fuse the skeleton data from multiple depth sensors and provide robust 3D skeleton data for avatar rigs.
- A hand tracking sensor subsystem was integrated to support a fully articulated avatar with natural hand gestures in VR.
- To enlarge the usable hand tracking area, A hand tracking system was proposed using multiple hand tracking devices (Leap Motion), and smooth the hand rig experience using a novel algorithm.
- The avatar system was improved in terms of eye and mouth movement rendering and refined the avatar rig algorithm to apply the fused skeleton and hand data to support communication and collaboration.

(b) The impact of the depth-sensor-based avatar system (full-body tracking with hand gestures) on communication behavior was investigated, and compared against a controller-based avatar system (partial-body tracking with limited hand gestures) in a single-user simulated interview application. I found that the depth-sensor-based avatar control system increased virtual body ownership and also improved the user experience. In addition, users rated their non-verbal behavior performance higher in the full-body depth-sensor-based avatar system.

(c) A shared virtual environment was implemented to investigate collaboration behavior using asymmetric avatar control systems. The effects of avatar expressiveness on co-presence, social presence, and interpersonal attraction were explored through a virtual charades game. A significantly higher social presence and interpersonal attraction was found when the participants interacted with users who were using the highly expressive avatar control system.

1.4 The Structure of the Thesis

In this thesis, an avatar system and two user studies are presented to address the questions in section 1.2 . The Human ethics committee of the University of Canterbury approved all the user studies. The thesis is structured as follows:

Chapter 2 addresses Research Question 1 and its sub-questions. In this chapter, the relevant research about the effect of avatar realism in avatar-mediated communication and collaboration is discussed. Then, I briefly go through a literature review about avatar related VR properties such as presence, sense of embodiment, and body ownership in the subsequent subsection. In the later subsections, a literature review of avatar control systems in terms of body and hand tracking, data fusion, and avatar rendering are provided. The last part of the chapter covers the current technology that can be used for the avatar system.

Chapter 3 addresses Research Question 2. Behavioral synchronization means that the pose and body movement of the user needs to be tracked in

real-time and then mapped to the virtual avatar. To answer question 2 and to get a contactless body tracking experience, three Kinect v2 cameras were used that capture the user and drive the avatar regardless of the user's orientation. System calibration and an adaptive data fusion method are described. Three different approaches are compared to fuse the data from three Kinects and compare against ground truth using an OptiTrack system. Two static poses and four movements were captured to compare the errors of each joint using the three fusion algorithms for the system evaluation. Results show that an adaptive weighting adjustment fusion method for combining skeleton data from the three Kinects according to the current facing direction performed best according to joint error, and variation curves are smoother than the other approaches.

Chapter 4 addresses Research Question 3. In chapter 3, the multiple Kinects solution is developed and evaluated, but it is not applied in VR. To answer question 3, a set-up using four Kinects was introduced for robust and accurate full-body 3D skeleton tracking together with Leap Motion integration into a Vive system. It was suggested that a calibration method to synchronize heterogeneous devices using a traditional checkerboard marker. New camera weighting methods were proposed and compared with previous approaches. The results showed that the improved adaptive weight calculation method proposed in this chapter could tackle several tracking issues. The results of the rotation fusion tests show that the system has good accuracy when compared to the Vive tracker. The Leap Motion data is integrated with the fused skeleton system, supporting natural hand gestures in VR interaction scenarios.

Chapter 5 addresses Research Question 4. Based on the system presented in chapter 4, a virtual interview was designed and implemented in a single-user simulated communication scenario to answer question 4. The participants went through an interview, which had two sessions, an interview session, and a route planning session with a sensor-based and controller-based avatar control system. The participants were encouraged to perform non-verbal behavior as well as a verbal cues to communicate with the virtual interviewer.

Specifically, the interview process was recorded in VR, together with all the verbal and non-verbal cues. Subjects then took a third-person view to evaluate their previous performance. It was found that a significantly higher virtual body ownership illusion and usability, as well as better non-verbal communication performance by participants in the depth-sensor-based experience compared to the controller-based experience.

Chapter 6 addresses Research Question 5. Some limitations were found after the user study described in chapter 5. The hand-tracking data sometimes switched between the fused Kinect system and the Leap Motion, and there was no finger data when the participants moved their hands outside the LMC tracked area. These issues were due to the limited tracking area of a single Leap Motion controller (LMC), and the user needed to put their hands in front of the HMD to avoid tracking loss. To answer question 5, a multi-LMC system was proposed. In this chapter, the configuration of the five-LMC system used on an Oculus Rift S is described. Then, there is a discussion of the shared-view calibration method for the system based on the Least-squares fitting (LSF) algorithm. To avoid incorrect tracking data from a single LMC interfering with the fusion result, a multi-LMC fusion algorithm based on two-level data evaluation was proposed, which consists of a prediction-based and a position-based evaluation method. Based on the evaluation result, the data from multiple LMCs with a Kalman Filter sensor fusion was combined. The experiment shows that the system can enlarge the hand tracking range to 202.16 degrees horizontally and 164.43 degrees vertically.

Chapter 7 addresses Research Question 6. In this chapter, a high level of non-verbal expressiveness avatar control system was presented, which combines with the work presented in chapters 5 and 6. The avatar control algorithm was optimized to make the avatar rig more smooth. Additionally, the eye and mouth movement rendering was improved, which is more realistic compared to the rendering in chapter 5. To answer question 6, a shared virtual environment using asymmetric avatar control systems was implemented. The effects of different levels of expressiveness of avatars on copresence, social presence, and interpersonal attraction were explored through a virtual

charades game. The participants took turns as word givers, and word guessers using the different avatar control systems, and they needed to collaborate to complete four sessions within the given time. It was found that a significantly higher social presence and interpersonal attraction when the participants interacted with users who were using the highly expressive avatar control system. Furthermore, participants had better task performance when they embodied a highly expressive avatar.

Following the studies, chapter 8 covers the discussions and conclusions arising from this thesis. Additionally, the future work that can be carried out to further the research that has been detailed in the thesis.

Chapter 2

Background

In the previous chapter, the motivations behind the research were introduced, and the related research questions were also covered. This chapter introduces the previous work about avatar-related communication and collaboration, VR properties, avatars, and the tracking methods that researchers have investigated.

2.1 Communication and collaboration behavior in VR

Communication is a basic skill for everyone [20]. Verbal and non-verbal behavior, which reflects social interaction, supports communication and collaboration. Non-verbal communication behavior is usually presented in a mutual conversation through face-to-face, video conferencing [136], or embodied avatar in VR [102, 29, 116, 32]. The VR system that supports social interaction requires replicating the user's appearance and behavior. Appearance realism can have an impact on the sense of presence. Kwon et al. [70] investigated the level of realism on anxiety in job interviews. They compared cartoon realistic, photo-realistic, and real people for the interview. They found that it provoked a greater sense of presence from a more graphic detailed virtual human.

The avatar in a virtual environment can lead to virtual body ownership illusion (VBOI). Induced VBOI can change the way users behave in the real world. Kilteni et al. [65] compared two different avatar appearances (casual

dark-skinned, formal light-skinned) in the virtual musical drumming application. Both two avatar appearances induced strong VBOI, but the participants have substantial behavioral and cognitive changes in a dark-skinned condition. Both conditions were using the same motion-capture suit, and they only focused on the avatar appearance. Few numbers of research focus on behavioral realism, which leads me to think of the question "Do the different levels of behavioral realism induce different VBOI or sense of agency?". "How non-verbal behavior realism with the embodied avatar in VR changes the user's communication behavior in the real world?"

The non-verbal cues delivered by the virtual characters in the collaborative virtual environment influence the efficiency of task performance [102], and the user's embodiment can lead to higher social presence ratings compared to face-to-face interactions [116]. Heidicker et al. [52] compared three different avatars (full-body avatar with idle animation, a full-body avatar with motion-controlled, and floating avatar with only head and hands motion-controlled) and solved a collaborative task in a user study. The results showed that motion-controlled avatars with full representation of the avatar body lead to an increased sense of presence. Motion-controlled avatars and avatars that have only head and hands visible produced an increased feeling of co-presence and behavioral interdependence. Other non-verbal behaviors such as facial expression and eye-tracking also affect social interaction. Garau et al. [41] explored the impact of avatar realism, either visual and behavior realism, with gaze control on the quality of communication. A simplistic or more realistic avatar were compared and gaze control was singled out. It was found that independent of head-tracking inferred eye animations could positively affect participants' responses to an immersive interaction. However, in the pre-condition, the avatar needs to be highly visually realistic, which means the avatar needs to be at a certain level of visual realism when the effect of non-verbal behavior is explored.

A mirror is usually used in the single-user scenario [65, 24, 75] to evaluate the VBOI and communication behavior such as non-verbal cue. The user can identify the consistency between their real sense and virtual representation

with the help of a virtual mirror. Gonzalez et al. [46] changed the reflection of the virtual mirror to make the participants experience synchronous or asynchronous reflection from their body movement. It was found that participants felt strong VBOI in a synchronous mirror reflection, which implied that the real-time non-verbal behavior reflected on the virtual avatar is essential. The avatar system needs accurate and robust tracking with low latency to make the user feel strong VBOI from the first person point of view. For those collaborative virtual scenarios, the user's social behavior and performance can be judged by another person in the shared virtual environment.

A shared virtual environment (SVE) is important for communication and collaboration between multiple users in different physical locations. Previous research on SVEs can be found [120, 113, 7, 6] and more detail about interaction in SVEs can be found in [108]. The quality of the SVE can impact the synchronous multi-user virtual experience if all the users do not perceive the same state of the VE. Pan and Steed [93] developed an SVE to explore the impact of self-avatars on trust and collaboration using virtual puzzles with the HTC Vive and Unity UNET system, which is widely used for supporting multi-user networking. Self-avatar, no avatar, and face-to-face conditions were compared. However, the avatar was only a visual representation, and the movement was from the controller, which was not reflected on the virtual hands. Smith and Neff [116] implemented an SVE for negotiating an apartment layout and placing model furniture on an apartment floor to explore the communication behavior in embodied avatars. Participants could only use limited hand gestures driven by the controllers for communication. Roth et al. [99] proposed a software architecture using four data layers to augment social interactions by integrating behavior tracking such as body, eye gaze, and facial expressions into the SVE. This software architecture was able to support social communication, but participants missed hand-gesture cues. The research mentioned above either missed the natural hand gesture interaction or used controllers to present limited gestures in the collaborative task, which led me to think about the question "Does the level of expressiveness of an avatar in terms of non-verbal behavior impact communication and

collaboration behavior?"

2.2 VR properties

The avatar realism in terms of form and behavior have an effect on the user's sense of presence and embodiment in the single-user application. Also, avatar realism can impact the co-presence, social presence, and mutual communication and collaboration behavior. In the following subsection, the VR properties that were explored in the user studies are summarized.

2.2.1 Presence

Avatar realism can affect the presence that the user can feel in the virtual environment. Therefore, what is presence? The concept of presence has many definitions and meanings, which can be defined in a variety of ways [38, 137]. According to Heeter [51] and Steuer [121], presence is often defined as the sensation of "being there" and it is an associated conscious state [114]. Nowak and Biocca [90] divided presence into three dimensions, including telepresence, co-presence, and social presence. According to Schroeder [107], presence can provide the feeling that the user is "there" inside the media (telepresence) or with other entities (co-presence).

The term telepresence is used to describe a user who feels immersed in the virtual environment represented by the medium [121]. Gerrig [44] and Minsky [82] propose that telepresence is the sensation of being in a mediated space which is different from the physical location of the body. The term co-presence first comes from the work of Goffman, who explained that users in the virtual world could perceive each other [31], which means co-presence refers to a psychological connection to another person [89]. As for social presence, it is the feeling of the user, which makes people feel connected with others through the telecommunication system, according to Rice [97], Short et al., [110] and Walther [135].

2.2.2 Sense of embodiment

Embodiment is the sense of feeling when users use avatars in the virtual environment and interact with virtual objects or characters. Kilteni et al. [66] summarized the definition of embodiment in multidiscipline use and applications from five different perspectives, such as philosophy [12, 81], cognitive neuroscience, psychology [8, 48], robotics [35, 134], and presence [9]. They also put forward the three sub-components of the sense of embodiment to measure it in an easy way, which is composed of the sense of self-location, agency, and body ownership. The sense of self-location refers to the spatial sensation inside a virtual body not in the virtual world, and the sense of agency means the user can control the virtual body in the VE.

Virtual body ownership illusion

Body ownership can be referred to as the feeling of manipulation over the virtual body in the VE world, which can be induced as shown through experimentation [11, 40, 129]. For example, with the discovery of the Rubber Hand Illusion (RHI) [17], Botvinick and Cohen showed a rubber hand could be incorporated into the body representation through the use of appropriate synchronous multisensory stimulation, which can reveal information about the perceptual process. A visible rubber hand and an occluded real hand were stroked at the same time to induce the feeling of ownership. Virtual body ownership illusion (VBOI) refers to a self-consciousness of one's own body [40, 63], which is a critical component to indicate the level of presence and sense of embodiment [66]. Yee and Bailenson found a Proteus effect that the virtual avatar's appearance and behavioral characteristic influenced the individual's behavior changes, and it depends on VBOI [141]. To enhance VBOI, Gonzalez-Franco et al. [46]. and Jung et al. [62, 61] studied the influence of real-time behavior of the avatar using a virtual mirror that results in a higher sense of body ownership. The visuomotor is a significant factor for virtual body ownership [12, 62]. The freedom of agency that refers to the sensation of controlling the virtual body has been considered as an essential

factor for VBOI.

2.3 Avatar systems

Avatars are necessary to convey roles, behaviors, and location. The behavior realism of embodied experience in VR usually requires a high-precision and real-time avatar rendering system. This helps to elicit immersive feelings in users of their bodies through controllable virtual avatars, including precise body movement tracking for hands, fingers, eyes, and even facial expressions. The following is a review of the previous work done on body and hand tracking.

2.3.1 Fully articulated body and hand tracking

Full-body tracking in VR can provide immersive experiences, enabling users to interact with the VE from the first-person point of view. In the following section, the three main solutions for the full body and hand tracking are summarized.

Wearable tracking Systems

Generally, in order to have a high-quality full-body tracking experience, the user needs to wear a motion-capture suit, such as with the OptiTrack[1] system used in [65][117][101]. These systems can provide high accuracy and low latency tracking, which is the primary solution for full-body tracking in VR. However, these systems also suffer several problems, such as retroreflective optical markers being visually indistinguishable [53] and interference from non-marker objects in the tracking area that can reflect infrared light. Besides expensive motion-capture systems, full-body tracking can also be implemented using consumer VR devices like HTC Vive with Vive trackers[2]. Caserman et al. [21] built a full-body tracking system using an HTC Vive HMD and Vive trackers attached to the wrists and ankles to drive an avatar model

[1] http://optitrack.com/
[2] https://www.vive.com/us/vive-tracker/

in VR. The position and orientation data from the HMD and the trackers are processed before feeding into an inverse kinematics (IK) system. The trade-off is between the tracking accuracy and the number of trackers. The tracking accuracy can be improved by attaching more tracking markers, which would make the user feel encumbered.

Full-body Tracking Using Depth Camera

Instead of using wearable devices for body tracking in VR, many researchers use depth cameras to track the body, which provides a natural interaction without wearing cumbersome devices. Sra and Schmandt [118] built a social system for multiple users where the Kinect was used for body tracking and the Oculus Rift[3] for tracking head rotation. This is the same approach adopted by Collingwoode-Williams et al. [24] and Czesak et al. [27]. However, the two devices' data are not in the same coordinate system, which could be an issue for consistency. Fountain and Smith [36] proposed a real-time ambient fusion method for a single Kinect and the HTC Vive. The calibration method in this system combined the most accurate data for a given body part amongst all systems, but still needed the data from controllers for error correction. The solutions mentioned above all used a single Kinect for body tracking, which restricts accurate movement capture of the user. A single Kinect cannot recognize which side of the body is facing the device, leading to the recognition issues when the user turns around.

Full-body Tracking Using Multiple Depth Cameras

To address the single Kinect tracking issue, some researchers suggested using multiple integrated Kinect devices. Müller et al. [86] placed six Kinect v2 sensors along a corridor with three Kinects on each side to assess gaits using 3D reconstructed models, and skeletons fused by the left- and right-side devices. Kaenchan et al. [64] placed three Kinect v1 devices at different angles and transformed the coordinate systems of two other Kinects into a reference Kinect. This method can handle the problem of occlusion to some extent,

[3]https://www.oculus.com/rift/oui-csl-rift-games=star-trek

but it does not consider the front side and back side recognition issues. Also, they used averaging to fuse the data, which could be unreliable when camera calibration errors are present. Kim et al. [68] built a 360-degree motion capture system with six Kinect v1 devices around the user and chose the Kinect that is in front of the user at every frame to fuse the skeletal data. In their later work [69], they proposed a method to use the right-left shoulder line as a pose vector to compare with the front vector for checking whether the user is facing the Kinect or not. They only use data from the three Kinects in front of the user, which is not strictly 360-degree motion capture, as they discard data from the other three devices every frame. Morato et al. [128] set up a four-Kinect system and fused data by inputting the mean of all Kinect measurements into a particle filter algorithm, using the mean value of the four Kinect measurements as the input is very sensitive to the user's orientation. If the user turns to a bad tracking direction, the data from the Kinect on the bad side would be extremely unstable and negatively impact the fusion accuracy.

Body and Hand Tracking Using Depth Cameras

A classical approach to hand tracking includes wearing gloves [28] or wearable cameras [67], but these are not natural to use for interaction. The Leap Motion is able to provide articulated hands based on the calculated depth information, and it also supports HMDs. Adam and Sreenath [26] fused the data from a Kinect and a Leap Motion device for a VR game. However, they only fused data pertaining to the palm's open or closed state. Morgado et al. [84] proposed a framework to make the Leap Motion and Kinect work in the same system for separate gesture detection. Amir et al. [1] built a VR first-person shooter game that also combined the Oculus Rift, Kinect, and Leap Motion. The customized postures are set for different interactions. The user needs to learn the mapping set to control the avatar, which is not the natural way to interact with the virtual environment.

2.3.2 Avatar control and rendering

An avatar is a virtual representation of a user and is driven by the user's movements in the virtual world [5]. An avatar system can provide an embodied experience [112], that users can interact with the virtual world through the avatar body from the first-person point of view through the virtual camera located at the virtual avatar's eye level. Early avatar control systems could not provide a complete embodied experience due to limited tracking technology (tracking area and accuracy), which led to reduced information channels such as body movement, hand gestures, and facial expressions. Current technology is still unable to capture and represent all non-verbal behavior from a single peripheral input device. To address the problem, representing highly expressive avatars for non-verbal behavior, integrating multiple sensors shows strong potential.

Body movement is the primary source of data for avatars. To get a high quality embodied experience, motion-capture suits are widely used in avatar-related research [65, 117, 101]. However, these are expensive and cumbersome, although providing high accuracy and potentially large tracking areas. In contrast, consumer VR devices such as the Oculus Rift or HTC Vive with spatial controllers are alternative solutions for tracking parts of the body. However, if more parts of the body want to be tracked, extra sensors are required. For example, most VR applications are based on this three-tracking-point (one HMD plus two controllers) solution, with only support "floating" avatars, such as in *Facebook spaces*[4], *VR Chat*[5], and *Mozilla Hubs*[6]. Extra trackers are required along with specific inverse kinematic software if the player needs a full-body experience [21]. Compared to the HMD and trackers solution, RGB-D camera-based body tracking is a contactless way that can provide more joint information. Relevant research can be found in [69]. Users can experience improved articulation control of their avatar using these approaches.

[4]https://www.facebook.com/spaces
[5]https://vrchat.com/
[6]https://hubs.mozilla.com/

Hand gestures are another essential data source, which can present important non-verbal information. The VR controllers can trigger specific gestures when certain buttons are pressed, but the remapping strategy is limited. To compensate for these constraints, a hand tracker, such as the LMC, can provide natural hand gestures without using any controller.

Other non-verbal cues for avatar control are eye gaze and facial expression. Roth et al. [103] used an RGB-D sensor to track facial expressions, and eye gaze then mapped the data on to an avatar. In later work [99], a system architecture was presented for the augmentation of social behaviors in multi-user environments. The avatar framework can present the non-verbal behavior such as facial expression and body posture, but it lacks hand gestures, which is not suitable for the hand gesture-based communication and collaboration task.

2.4 Tracking methods in VR

Positional tracking is an essential part of the VR experience to achieve immersion and presence. Since the user should see the computer-generated environment using an HMD, head-tracking should be supported to render the scene according to the user's head movement, and this gives the minimal VR experience. For more advanced VR experiences, additional tracking devices can help users to experience the realistic sensation between real-world and virtual environments (VEs) by tracking gestures, facial expressions and other movements of the user. In this section, the solution that can support natural behavior of the user in VR is summarized.

2.4.1 Head tracking

In the early stage, many solutions were used for head tracking such as a mechanical tracking system, an electromagnetic tracking system, an acoustic tracking system, and an Inertial Measurement Units(IMU) system. Due to

the accuracy and latency, the current VR system for head tracking generally applies optical and computer vision technology.

Outside-in and Inside-out tracking

The mainstream VR hardware products such as Oculus Rift and PlayStation VR pro[7] use outside-in technology to track headset and accessories, which all need external sensors as Figure 2.1. High accuracy and low time-latency are the main advantages of these VR systems due to the external sensors, but this also causes some problems. Occlusion is the main problem. Also, they do not scale as well as inside-out, since each tracker element adds to the complexity of finding the captured images. Another limitation is that the users are tracked as long as they're in the field of view. If they step outside of the tracking area, it can be particularly problematic.

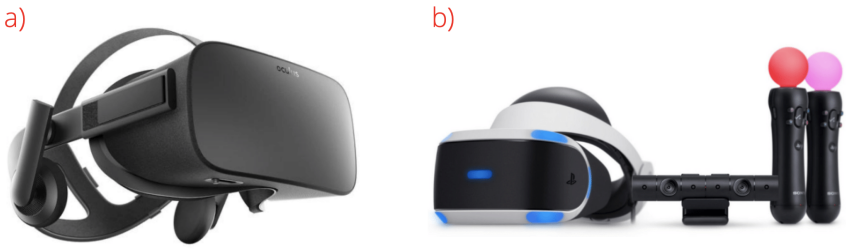


FIGURE 2.1: Outside-in tracking, a) Oculus Rift, b) PS VR

Inside-out tracking is usually applied to headsets like Hololens[8] and Oculus Rift S[9] as Figure 2.2. The built-in cameras are using markerless inside-out tracking, which determines how its position is changing with the external environment. Mobility is the main advantage of inside-out tracking when compared to outside-in tracking. HTC Vive[10] is using the external lighthouse as markers, which is marker-based inside-out tracking.

[7]https://www.playstation.com/en-nz/explore/playstation-vr/
[8]https://www.microsoft.com/en-nz/hololens
[9]https://www.oculus.com/rift-s/
[10]https://www.vive.com/nz/product/vive

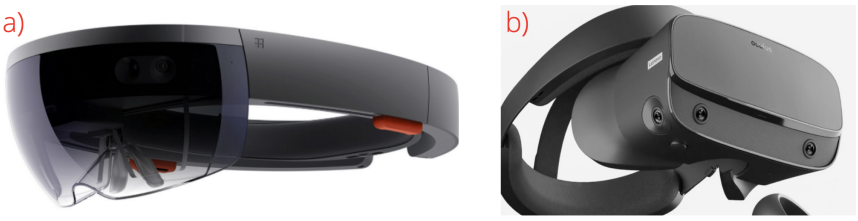


FIGURE 2.2: Inside-out tracking, a) Hololens, b) Acer Mixed Reality Headset

2.4.2 Body tracking

The embodied experience in VR could be improved by using real-time tracking data from the user. The motion capture system is used a lot for high tracking accuracy. This system normally consists of a set of markers that can reflect infrared light and cameras that can capture markers for calculating the position of the target. As it uses optical tracking technology, a fast transmit rate is the main advantage that minimizes the time latency issue. For example, OptiTrack[11] passive tracking (Figure2.3) uses reflective markers on the tracked person or object and uses dedicated cameras.

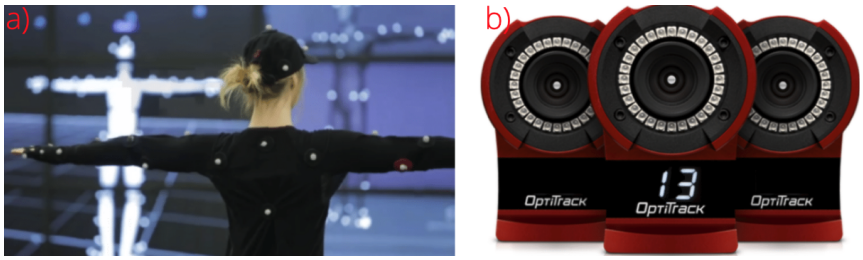


FIGURE 2.3: OptiTrack passive tracking: a) Tracking suit with reflective Markers, b) Cameras

Some other tracking systems use cameras to detect active LED makers on the person, such as Impulse X2E[12] motion capture system made by PhaseSpace (Figure 2.4). Users can be captured in real-time once they wear the vest, gloves, hat, guns with LED lights, and body movement can be recorded and analyzed by the matched software. The optical tracking systems mentioned above also have some disadvantages, such as occlusion issues. If the object obscures the

[11] https://optitrack.com/
[12] http://www.phasespace.com/impulse-motion-capture.html

line of sight between a camera and markers, the tracking would be affected. The latency could be another issue when the number of markers increased.

FIGURE 2.4: Impulse X2E motion capture system

As mentioned above, OptiTrack and PhaseSpace are motion capture systems, but the user needs to wear suits or markers, which is inconvenient. Microsoft Kinect v2[13] can track the skeletons of multiple users using an RGB-D camera as Figure 2.5. The sensor provides color information as well as the estimated depth for each pixel. However, the user has to face the device, because it cannot recognize which side of the user is facing the device.

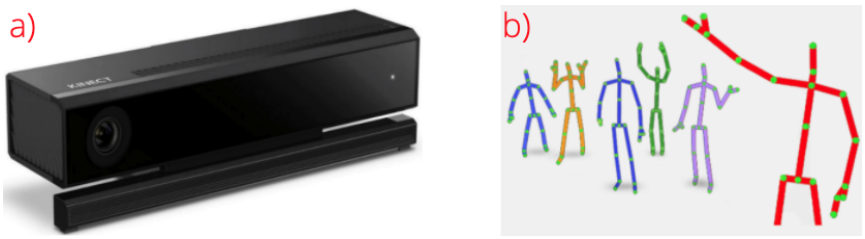


FIGURE 2.5: RGB-D camera, a) Kinect V2 sensor, b) Skeleton data

2.4.3 Hand tracking

Interacting with virtual objects is an essential component of VR experience. A controller can be used to pick or throw, but it is not the best choice for operations such as pinching and fingertip manipulation. Utilizing wearable

[13]https://developer.microsoft.com/en-us/windows/kinect

gloves, we can get the movement of fingers and hands. For example, VMG Lite Data Glove[14] (Figure 2.6a) can provide five high accuracy joint angle measures sensors, and transform finger and hand motion into real-time data. CyberGlove II[15] (Figure 2.6b) has 18-22 sensors inside each glove, three flexion sensors per finger, four abduction sensors, a palm-arch sensor, and sensors to measure wrist flexion and abduction. Each sensor is extremely thin and flexible, being virtually undetectable in the lightweight elastic glove.

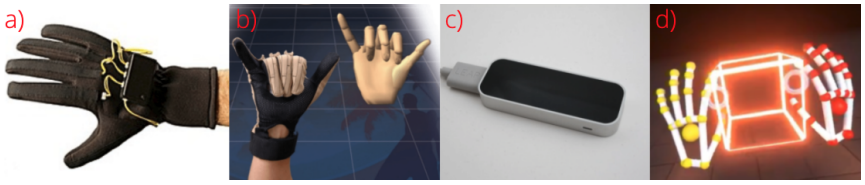


FIGURE 2.6: Hand tracking and gesture recognition, a) VMG Lite Data Glove, b) CyberGlove II, c) Leap Motion, d) Virtual hands rendered through Leap Motion sensor

Alternatives to the glove for hand gesture, computer vision-based devices such as Leap Motion[16] and DepthSense 525[17] Camera manufactured by softkinetic are depth cameras which can provide natural hand gesture and fingertip operations (Figure 2.6c and 2.6d). Leap Motion is compatible with most platforms and VR headsets, but it has a limited tracking range for the hands.

2.4.4 Eye tracking

Eye movement is one of the most natural ways we interact with the world. We normally need to gaze at the object before selecting it, which is common in the real world but is difficult in VR. Eye-tracking devices such as Pupil Lab[18] and FOVE[19] provide eye contact data and apply it to VR for multiple users' communication and object selection. The principle here is to detect the reflection of IR light on the eye. Pupil labs provides an add-on for Vive

[14]https://www.vrealities.com/products/data-gloves/dg5
[15]http://www.cyberglovesystems.com/cyberglove-ii/
[16]https://www.leapmotion.com/
[17]https://www.sony-depthsensing.com/
[18]https://pupil-labs.com/
[19]https://www.getfove.com/

and Oculus (Figure2.7a). FOVE is a VR headset with a built-in eye-tracking module(Figure2.7b).

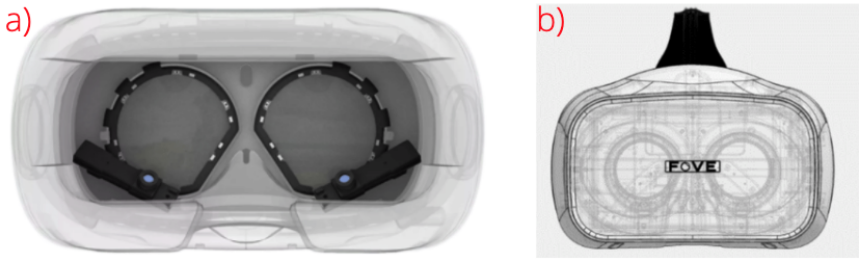


FIGURE 2.7: Eye tracking device for VR, a) Pupil Lab, b) FOVE

2.4.5 Facial expression tracking

Emotion is an expression of mental activity that can be detected by facial expressions [32]. Social VR provides a platform for multiple players to communicate, where rich facial expressions of the avatar can enhance the presence of the user as it reflects the player's true feelings [91], which is essential, especially in face to face scenarios. MASK[20] is a facial expression tracking device suite for mainstream VR headsets as it can be installed around the foam contact point of the headset. It has eight electrodes to detect signal waves produced by the facial muscle movement and convert the different signal waves into relevant emotion, which is based on neuro-VR technology (Figure 2.8a). BinaryVR[21] is another real-time facial tracking device that can attach to mainstream headsets. The camera can detect mouth movement and map to the facial expression of the avatar (Figure 2.8b).

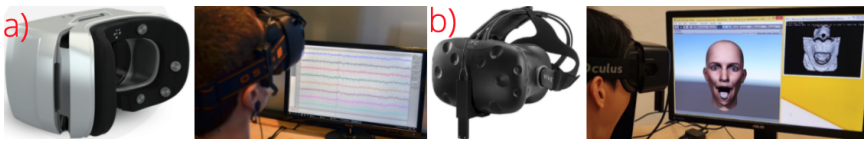


FIGURE 2.8: Facial expression tracking device for VR, a) MASK, b) Binary VR

[20]https://www.theverge.com/2017/4/13/15251616/mindmaze-mask-vr-face-expression-reading-sensors
[21]http://www.binaryvr.com/

2.5 Conclusion

In this chapter, the first thing that was covered was avatar realism's impact on communication and collaboration behavior. Then a brief discussion of VR properties that may affect the user during the embodied VR experience. The avatar system and control solutions in the previous work was summarized, and the current technology that can be used to build the avatar system. From the previous research, it is known that avatar realism in terms of non-verbal behavior can impact the user's presence, sense of body ownership, communication, and collaboration behavior. However, expressive avatar systems (integrating non-verbal behavior, such as body movement, hand gesture, facial expressions, and eye gaze) are limited in current immersive systems due to sensory technologies. In the next chapters, the highly expressive avatar control system in VR, and the effect on user's communication and collaboration behavior will be explored.

Chapter 3

Robust 3D Skeleton

An essential aspect of immersive social experiences in VR is the user's representation, or avatar. Full-body tracking data can be used as the primary data source for an avatar control system. The Microsoft Kinect v2 sensor can provide skeleton data of users in real-time. However, due to occlusion issues and front/back ambiguity errors, one Kinect is not always reliable enough. In this chapter, I present work to provide robust, real-time tracking using multiple Kinect v2 cameras. An adaptive data fusion method is described that constructs a high-quality 3D skeleton that can be used to drive the avatar regardless of the user's orientation. This work was presented as a paper [138] in the 10th International Conference on Virtual Worlds and Games for Serious Applications (VS Games 2018), which was held in Würzburg, Germany from 5th to 7th November 2018.

3.1 Introduction

Microsoft Kinect v2[1] can track the skeletons of multiple users using an RGB-D camera. Zhang [144] explains that the segmentation process from depth images uses per-pixel classification, and each pixel is evaluated separately to go through the pipeline (Depth image->Inferred body parts->Hypothesized joints->Tracked skeleton) to get skeleton data. Although Kinect v2 provides improved accuracy, field of view, number of joints, and number of people detected compared to Kinect v1, according to Samir et al. [106], the issue of

[1]https://developer.microsoft.com/en-us/windows/kinect/

occlusion still exists. The state of a joint is inferred, not tracked, when it is occluded, which can distort the skeleton. Also, the Kinect v2 is not able to recognize which side of a person (front/back) is facing it, which decreases the pose accuracy and motion accordingly as the person moves or turns. To solve the problem, I investigated if a multiple Kinect solution can address this issue by integrating the data from each Kinect to optimize the accuracy of the overall skeleton and correctly detect the direction the user is facing.

In this chapter, the set up is described for a tracking system that adopts a client-server approach, where each client is connected to one Kinect and sends skeleton data (25 joints) through an Open Sound Control (OSC) message to a server machine for fusion and smoothing. In the server machine, UniOSC [130] plugin was used for Unity to handle OSC messages and created a novel adaptive 3D skeleton data fusion algorithm to process data from each client. This algorithm consists of real-time facing direction detection, left- and right-side swapping (LRS) of joint info, adaptive weighting adjustment for each camera, and weighted averaging for each joint from the three Kinects. A double exponential smoothing filter was utilized to optimize skeleton data before sending it to the server machine, in order to reduce jitter and provide smoothing.

The remainder of this chapter is organized as follows. The system setup and configuration, camera calibration, data fusion, and smoothing filters are presented in Section 3.2. In Section 3.3, the design of the experiment is described to evaluate the proposed algorithms by comparing the error of each fused joint with ground truth data from an OptiTrack[2] system. In Section 3.4, results and discuss future work are summarized.

3.2 System

The system uses a client-server approach with three Kinect v2 devices directly connected to three client PCs through USB 3.0 (Figure 3.1). The three client machines retrieve skeleton data using the SDK and send it to the server PC

[2]https://optitrack.com

through local Ethernet. All the skeletal data processing is done on the server before streaming it to Unity for visualization. Three Kinects are placed around a 2.4m-radius circle on the tripods, at the height of 1.7m, 120° from each other, dividing the detection area into three regions. Since the user is standing on a 20cm raised floor, the height of the tripods is 1.5m from the raised floor. In order to evaluate my three-Kinect system, I installed six OptiTrack flex13 cameras on the frame of the cage 2m above the cage floor to capture ground truth data.

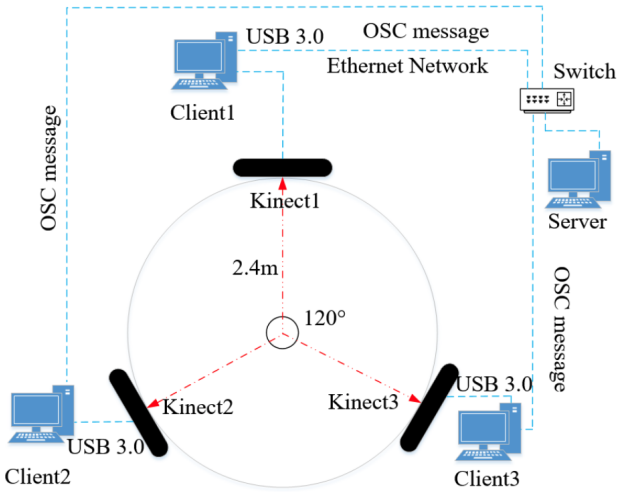


FIGURE 3.1: Multiple Kinects setup

.

3.2.1 Calibration and Pre-processing

As the skeletal data from each Kinect uses a coordinate system relative to the Kinect itself, we need to transform the three separate coordinate systems into a reference world coordinate system (Equation 3.1). Generally, the cameras can be calibrated based on computer vision algorithms by using a "chessboard" pattern [142]. A chessboard is placed on the cage floor to be captured by three Kinect cameras at the same time. In this case, the projection matrix from each Kinect to the chessboard world coordinate could be calculated using the OpenCV SolvePnP function [92]. This step can be done one time, and the resulting matrices are loaded when the application starts. Then, all the three

skeletal data sets are rendered in the same coordinate system and implement the data fusion in the next step.

$$\begin{bmatrix} X_w \\ Y_w \\ Z_w \end{bmatrix} = M \begin{bmatrix} X_k \\ Y_k \\ Z_k \end{bmatrix} \tag{3.1}$$

Here, (X_k, Y_k, Z_k) is the coordinate of each Kinect, M is transform matrix, which contains rotation matrix R and translation matrix T, (X_w, Y_w, Z_w) is the related position in the world coordinate system. Furthermore, in order to ensure the skeleton data of the Kinects is stable, double exponential smoothing [60] is used recommend by Microsoft for jitter reduction and smoothing before sending the data to the server machine.

As a control, the skeletal data from the OptiTrack also needs to be transformed into the same world coordinate system. The OptiTrack has its calibration wand with three reflective markers in an L-shaped tool. When placed on the flat surface of the cage, this calibration tool indicates the x-axis and z-axis, and the y-axis is in the upward direction. This calibration method is implemented by placing the three reflective markers on the corners of the chessboard on the cage floor. In this case, the coordinate system of the OptiTrack has been manually set to the same chessboard world coordinate system as the Kinect cameras.

3.2.2 Facing Direction Calculation

As previously discussed, the Kinect cannot recognize the front and back sides of the user, so it is necessary to estimate the facing direction for each frame in order to fuse data correctly. A similar method is adopted as Kim et al. [68] and Kwon [69]. The main difference is that joints from the right and left sides of the whole body are used, instead of just the shoulders to make a body vector (BV), and then use the cross product with the fused facing direction to determine whether we are viewing the front or back from each Kinect. If the right and left shoulder joint were only used as the BV, it would not be reliable,

as one joint may occlude another one, e.g., when the user faces 90° from the Kinect (see, for example, Figure 3.4a).

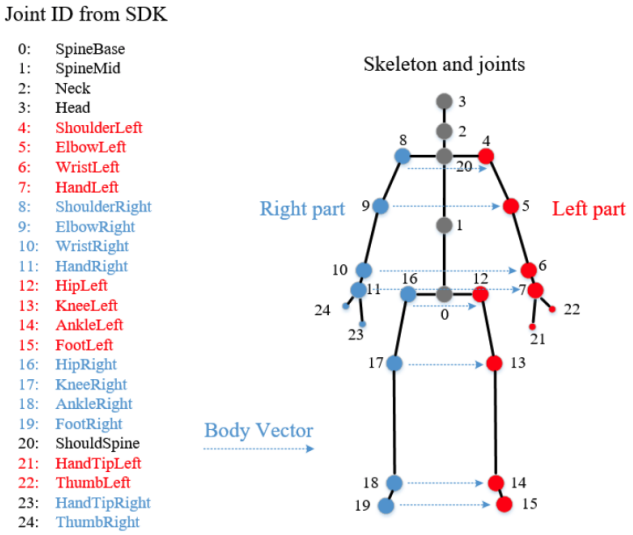


FIGURE 3.2: Joints and body pairs

When the session starts, the user stands in the cage center and faces in any direction. I collect the three sets of skeleton data and the skeleton whose joint states are all reported as "tracked" by the API will be used as the default data source to calculate initial facing direction using the BV described above, and Equation 3.2:

$$V_{initialFD} = V_B \times V_{Up} \tag{3.2}$$

where V_B is one pair of the eight BVs shown as blue arrows in Figure 3.2, V_{Up} is the unit upward vector and V_{FD} is the facing direction vector shown in Figure 3.3.

The facing direction is then updated once per frame and compare it with the BV from each Kinect using the cross product to check whether the data is from the front or backside of the user, which is a crucial step for later LRS processing. If the result is less than zero, then the data from this Kinect is the backside of the user, or vice versa, as shown in Figure 3.3.

The rule for choosing a BV for each Kinect is according to the joint tracking state reported by the SDK. For each joint, there are three values: tracked,

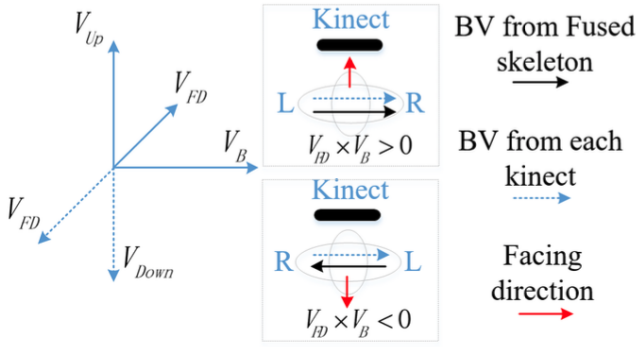


FIGURE 3.3: Facing direction detection for LRS

estimated, and not tracked, and a different weighting to each state is assigned, as shown in Equation 3.3.

$$\lambda_i^j = \begin{Bmatrix} 0, & NotTracked \\ 0.5, & Estimated \\ 1, & Tracked \end{Bmatrix} \tag{3.3}$$

where $i \in \{1, 2, 3\}$ is the i^{th} Kinect and $j \in \{1, 2, ..., 24\}$ is the j^{th} joint. For every frame, I calculate the product of eight pairs, for example $\lambda_i^8 \cdot \lambda_i^4, ..., \lambda_i^{19} \cdot \lambda_i^{15}$ and choose the BV if the value equals 1.

3.2.3 Skeleton Data Fusion

Data fusion can be carried out once the facing direction is calculated and the relationship with each Kinect. In this system, the LRS method is adopted to process the skeleton data from the back of the user. When the user is facing away from the Kinect, the device assumes the user is facing towards it, as it cannot distinguish front and back. All postures and gestures would be classified reversely, so the left side of the body would be incorrectly recognized as the right, and vice versa. In this system, the joints are grouped as either being on the left or right of the body (Figure 3.2). When the system determines that the data from a Kinect is from the back of the user, the LRS function will be called to swap the classification of each joint from left to right (and right to left), then use a weighted average method to fuse the three skeletons. All the

data calculation and manipulation are updated every frame. The weighted average method is shown in Equation 3.4:

$$\overline{P_j} = \frac{\sum_{i=1}^{3} \lambda_i^j \cdot P_i^j}{\sum_{i=1}^{3} \lambda_i^j}, j \in \{0, 1, ..., 24\} \tag{3.4}$$

where $\overline{P_j}$ is the fused position of the j^{th} joint and P_i^j is the position of the j^{th} joint from the i^{th} Kinect. Please note that this method is called "Condition 2" in the later evaluation section.

Condition 2 is considered the integration of the facing direction calculation, LRS, and weighted averaging. Although it is sufficient for comparing with just using one Kinect, it still does not consider the weighting of each Kinect. When we stand in front of a Kinect within a certain angle, the data is much more reliable and stable. Therefore, another method is adopted to fuse the skeletons, which can automatically assign weights for the values from each Kinect according to the current facing direction.

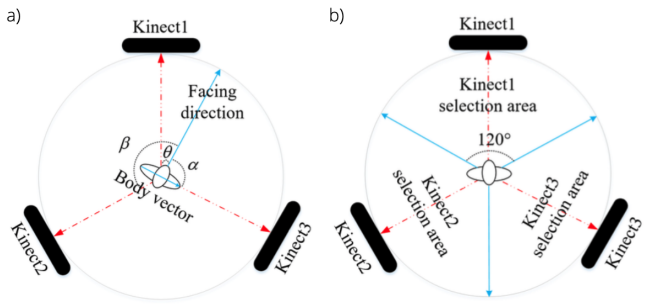


FIGURE 3.4: Angle calculation. a) Angle calculation and weighting assignment, b) Kinect selection area for Condition 1

The circular area is divided into 360 degrees (Figure 3.4b), and calculate the positions of each of the three Kinects in the world coordinate system using a calibration step. The angle between the facing direction and each Kinect direction is calculated according to Equation 3.5 every frame (Figure 3.4a), and assign a weighting for joints from each Kinect automatically according to Table 3.1.

$$\alpha = \arccos\left(\frac{\vec{F}\cdot\vec{K}}{|\vec{F}|\cdot|\vec{K}|}\right) \tag{3.5}$$

Here, $\vec{F}$ and $\vec{K}$ are vectors for the user's facing direction and the Kinect, respectively.

TABLE 3.1: Weightings

Angle	$0° \sim 90°$	$90° \sim 180°$	$180° \sim 270°$	$270° \sim 360°$
Weight range	$90 \sim 0$	$0 \sim 90$	$90 \sim 0$	$0 \sim 90$

When the angle increases by $1°$, the weighting for this Kinect decreases by one from the last state. Therefore, the weight for each Kinect is changing continuously and smoothly when the user moves around in this multiple-Kinect system. The modified fusion method is described by Equation 3.6.

$$\overline{P_j} = \frac{\sum_{i=1}^{3} \lambda_i^j \cdot \beta_i \cdot P_i^j}{\sum_{i=1}^{3} \lambda_i^j \cdot \beta_i}, j \in \{0, 1, ..., 24\} \tag{3.6}$$

where β_i is the weighting for the i^{th} Kinect. Please note that this method is called "Condition 3" in the later evaluation section.

3.3 Evaluation

The skeleton data retrieved from this system will be used to drive the movement of a virtual avatar. Hence, accuracy and error deviation is significant for this system. Several poses and movements were captured and recorded for multiple Kinects and OptiTrack system for evaluation (Figure 3.5).

3.3.1 Three-method Comparison

Kim et al. [68] and Kwon et al. [69] compared six Kinects to three Kinects. When the number is three, they only use the Kinect which is in front of the user. I call this 'Condition 1" (Kim and Kwon's method) and divide the Kinect

FIGURE 3.5: Evaluation setup and fused skeleton in condition 3

data source selection area. The angle between the facing direction and the Kinect direction is calculated every frame to choose the fused data from the Kinect, as shown in Figure 3.4b. All three conditions can be summarized as follows:

- Condition 1 (Kim and Kwon's method): Calculate the angle between the facing direction and each Kinect every frame, and choose one Kinect as the data source.
- Condition 2: Calculate and update the facing direction compared to the BV of each Kinect to recognize the front or backside of the user, applying the LRS method to process the data when the user's back is recognized for the relevant Kinect. Use a weighted average fusion method for data fusion.
- Condition 3 (My method): Similar to Condition 2, but calculate the angle between the facing direction and each Kinect direction and automatically assign weights for each Kinect according to this angle.

Two static poses, T-Pose and Squat, and four movements, Arm flapping, Walking, Upper-body rotation, and Crouching were designed. For 360° evaluation of the system, the experimental space was divided into eight regions with 45° for each region. Each static pose and movement was recorded through the Kinects and OptiTrack at each of the eight angles. To evaluate the quality of

tracking data, the Euclidean distance between Kinect joints and the relevant OptiTrack markers was calculated. As Kinect hand gesture recognition is not stable, 21 of the 25 joints (0-20) are selected from the Kinect (joint IDs are shown in Figure 3.2).

Eight groups of data for each condition were collected with the OptiTack, and averaged the data, which are presented in Table 3.2, with error curves in Figure 3.6.

TABLE 3.2: Average error (cm)

	Condition 1	Condition 2	Condition 3
T-pose	11.035	10.12	8.71
Squat	12.01	11.19	9.48
Arms Flapping	14.41	12.72	9.83
Walking	15.48	13.42	11.01
Upper-body rotation	15.42	13.80	11.97
Crouching	10.25	8.78	6.77

Repeated measures ANOVA test was used to compare three conditions differences for two static poses and four movements. The statistic results for T-pose ($F(2, 42) = 4.57, p = 0.016$), Squat ($F(2, 42) = 4.58, p = 0.016$), Arms Flapping ($F(2, 34) = 4.164, p = 0.024$), Walking ($F(2, 34) = 3.616, p = 0.038$), Crouching ($F(2, 34) = 5.793, p = 0.007$) show that there are significant difference between the three conditions. But the result for Upper-body rotation ($F(2, 34) = 3.214, p > 0.05$) shows that there is no difference between three conditions, which is because that the head joint selected for upper-body rotation have similar error under three conditions during the movement.

From the first two rows of Table 3.2, it is clear that the average errors for Condition 3 are lower than for Condition 1 and Condition 2, because Condition 3 integrates the angle calculation for Kinect weighting into Condition 2. From Figure 3.6, we can see that the curve for Condition 3 is smoother than for Condition 2 and Condition 1, which is due to continuous Kinect weighting distribution.

From the pose graphs, it is evident that Condition 3 is significantly more stable and smooth. The jitter and variation of Condition 1 are evident because

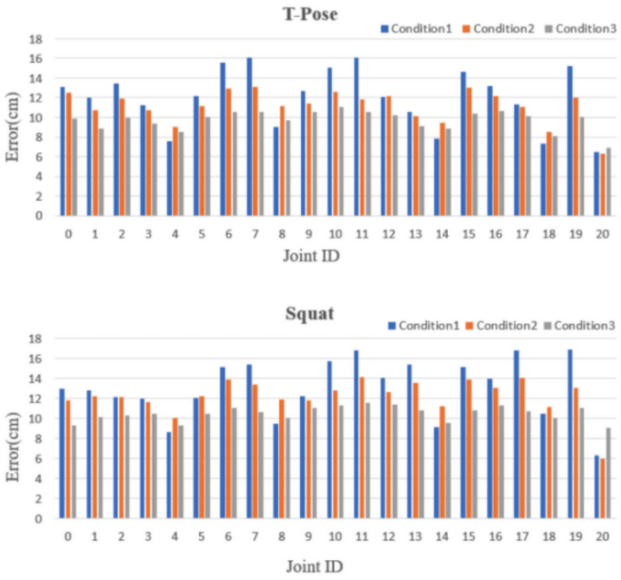


FIGURE 3.6: Two static pose: T-pose and Squat

the data source changes abruptly when the user's facing direction changes, so the data is unstable, especially when the user is facing a boundary between two Kinects. The average error of the joints of the wrist (6, 10), Hand (7, 11), Foot (15, 19) are apparent, which is also because the data source is changing rapidly, as data from this part of the body is quite different from different Kinects. The results for joints in the shoulder (4, 8, 20) and Ankle (14, 18) are much more stable in all three conditions, as these parts are in the middle of the user, and two poses have little impact on them no matter which direction the user is facing.

As these issues in static poses are apparent, movements comparison were considered to evaluate the three conditions for the relevant joints. In the movement evaluation, one specific joint was chosen for each motion, such as the right wrist for Arm flapping, the right knee for Walking, the head for Upper-body rotation, and the shoulder center for Crouching. The subject stood in the center of the detection area and executed the relative movement in the eight angles for the three conditions. The results are shown in Table 3.2, and Figure 3.7 shows the average errors of the specific joints over a continuous period.

As can be seen in the last four rows of Table 3.2, the average errors are

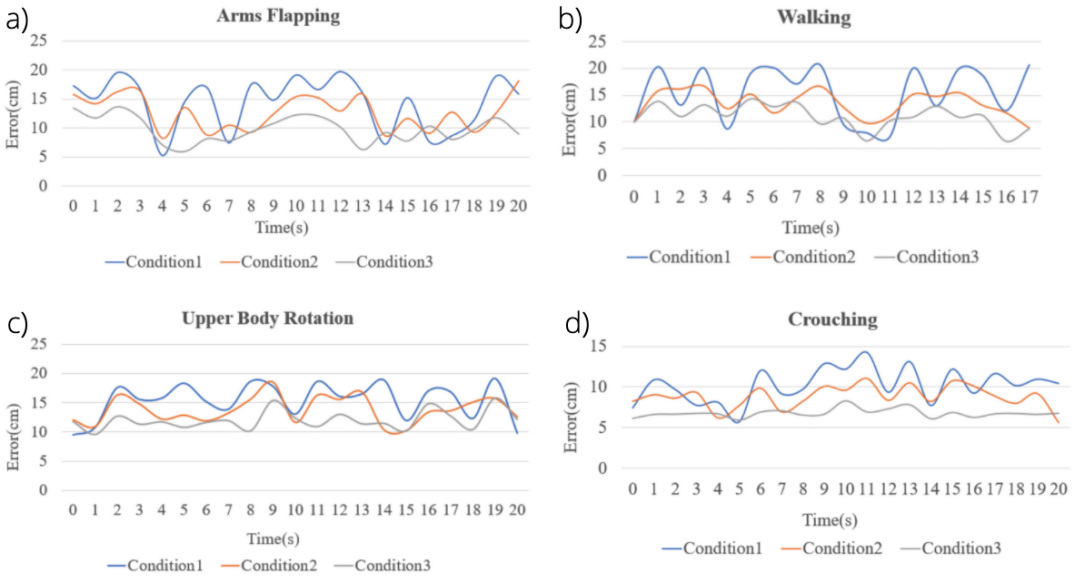


FIGURE 3.7: Four movements: a) Arms flapping, b) Walking, c) Upper body rotation, d) Crouching

similar to the static pose data, and in the variation graphs (Figure 3.7), the curves for Condition 3 are smoother than the other two fusion methods. The curve for Condition 1 in all four movements is the worst one, so it can be concluded that choosing the main Kinect according to the angle is not reliable, as the data source may suddenly jump to the next Kinect at the boundary areas. The results from Condition 2 are acceptable, as this method includes the front and backside recognition issue of Kinect, but without considering the weighting issue for each Kinect. Condition 3 (my method) integrates weighting assignment with data fusion methods from Condition 2, which are very reliable for full-body tracking. From Table 3.2, we can see that average errors for each joint in the three conditions are all around 10 cm; this is caused by errors from the Kinect SDK for skeleton recognition and body thickness caused by the marker location on the surface of the user.

3.4 Conclusion and future work

In this chapter, a multiple-Kinect system is setup for robust and high-quality full-body 3D skeleton tracking. System evaluation was carried out by comparing the Euclidean distance errors between three Kinect methods and an

OptiTrack system. The results show that the proposed method of adaptive weighting adjustment for three Kinects according to the facing direction of the user, and left- and right-swap (LRS) performed better than the other two methods reported in the literature. In the next chapter, hand-posture capture will be added to the system, further fusing the data with the current method of multi-Kinect capture described here. The goal is to provide a robust capture mechanism for immersive VR experiences.

Chapter 4

Towards Greater Avatar Articulation in VR

An increasing number of VR applications now use virtual avatars to represent the user in VEs. To fully control these virtual avatars, movement-tracking technology is required. In Chapter 3, a system was presented to provide contactless body tracking, but the system had no integrated support for tracking hand gestures. In this chapter, I investigate further based on the work in Chapter 3 for accurate full-body movement to include hand and finger tracking. This provides users with the possibility of using natural gestures to interact in the VE. In particular, I improve on Chapter 3 in the following five aspects. I have, (1) extended the calibration procedure to eliminate the tracking offsets between the RGB and depth cameras, (2) optimized facing-direction detection to improve the stability of data fusion, (3) implemented two new weighting methods for the depth data fusion of multiple cameras, (4) added the ability to fuse joint-rotation data, and (5) integrated a short-range depth camera for finger tracking. The system was evaluated empirically and show that the new methods improve the previous work in terms of tracking accuracy, and notably reduce the coupled hand-lifting phenomenon. This work was published as a full paper [140] in the Journal of Entertainment Computing Volume 31, August 2019.

4.1 Introduction

Applications like *Facebook Spaces*[1] and *VR Chat*[2] that represent users as avatars are becoming more and more popular. However, they are often unable to provide full-body avatars that are accurately controlled by users.

Normally, head-mounted display (HMD) devices and hand-held controllers are mainly used for head and hand tracking. For avatar control, the body motion still needs to be computed using Inverse Kinematics (IK), such as the approach described by Aristidou et al. [3]. However, this approach sometimes provides unrealistic results when the user assumes a complicated posture or makes a complex gesture. If more accurate full-body tracking is required, the user has to wear other tracking devices such as Vive trackers on the feet. Alternatively, the user can use expensive and cumbersome marker-based tracking systems that require them to wear tracking suits, such as OptiTrack motion capture system[3]. In terms of hand and finger tracking, consumer devices like Oculus touch and the HTC Vive controller can be of limited use due to the hand gesture and buttons mapping. Tracking gloves, on the other hand, could be used, but are generally expensive and again require that the user wear additional devices.

Consumer depth cameras are a cheap alternative to expensive and cumbersome marker-based motion capture systems used for body posture or hand gesture recognition. However, RGB-D cameras such as Kinect v2 have occlusion issues, and front/back ambiguity errors. To solve these, a multi Kinect v2 system was proposed in Chapter 3. The skeleton data, which comes from three Kinects around the user, were fused with a weighted average method, thereby allowing free movement of the user in the tracking area. Although the Kinect-based body tracking performed well, hand tracking was not accurate or stable due to few recognized finger joints . The Leap Motion device can provide fully articulated finger tracking and recognize natural hand gestures using a short-range depth camera. Hence, in order to allow the user to have

[1] https://www.facebook.com/spaces
[2] https://vrchat.com/
[3] https://optitrack.com/

a fully controlled VR experience, a fused Kinect tracking system needs to be integrated with a Leap Motion and VR system.

In this chapter, two improved methods based on weight factor are proposed in Chapter 3 to refine the tracking quality. The remainder of this chapter will cover the system set-up, camera calibration, and data fusion algorithms. Then, two new camera-weighting methods will be compared with previous ones and analyze the differences. A fused-skeleton system and Leap Motion integration can be seen in section 4.4. Finally, in section 4.5, the results and discuss future work are summarized.

4.2 System

In this section, an articulated full-body tracking system will be introduced, including finger movements, for a VR user. For body tracking, four Kinect v2s are used to enlarge the tracking area without directional hindrance. To manage the sensor data from the multiple Kinects, a client-server model was designed where the four client machines retrieve skeleton data using the Kinect v2 SDK and send it to the server PC through local Ethernet. All the necessary skeletal data processing is done on the server before streaming it to Unity for visualization. The four Kinects were installed on tripods placed at the corners of a square with 3.2m length and width (see Figure 4.1). Due to the Infrared (IR) interference problem between HMD and Kinect, the Kinect's hight was adjusted from 1.7m to 1.2m to avoid the problematic situation where the HMD faces a Kinect directly while the user looks around. The HTC Vive Pro[4] system was installed in the same tracking area with two lighthouses placed by the side, and the Leap Motion was attached at the center of the HMD for articulating finger movement.

[4]https://www.vive.com/nz/product/vive-pro/

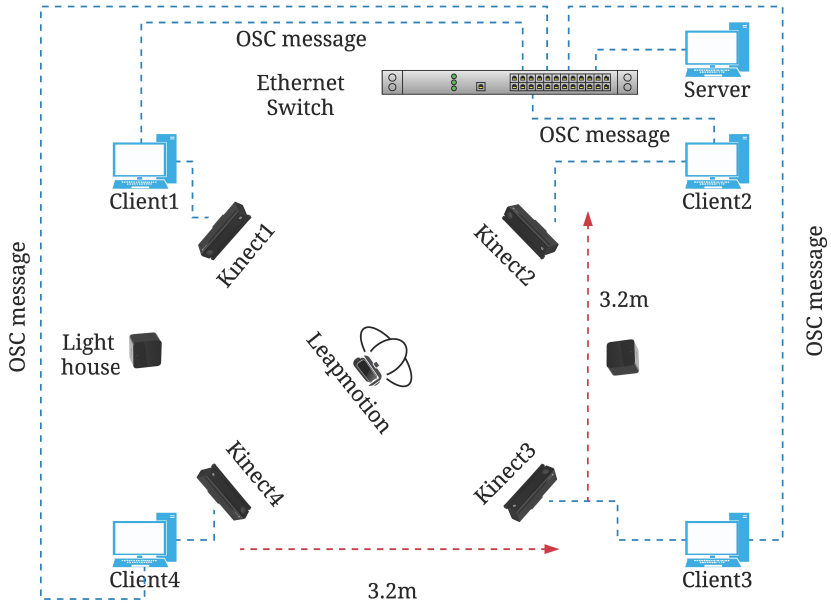


FIGURE 4.1: Multiple Kinects setup

.

4.2.1 Calibration

A calibration process has to be conducted to set up a common coordinate system for fusing the data between the four Kinects and the Leap Motion. A checkerboard was used as the world-coordinate origin for the Kinects. The Leap Motion is attached in the center of the HMD, as recommended by the official tutorial [72]. The local coordinates of each Kinect was converted to world coordinates using the checkerboard as a first step, and the calculated transformation matrix for the world to Vive system for the HMD. The coordinate system transformation pipeline was conducted in the following order: *Kinect* $\rightarrow$ *Checkerboard* $\rightarrow$ *HTC Vive system* $\rightarrow$ *HMD* $\leftarrow$ *Leap Motion*.

Kinect to Checkerboard

Once a checkerboard in the center of the tracking area was placed, the projection matrix between each Kinect to the checkerboard was calculated for the world coordinate origin using the OpenCV SolvePnP function [92] as shown in Figure 4.2a.

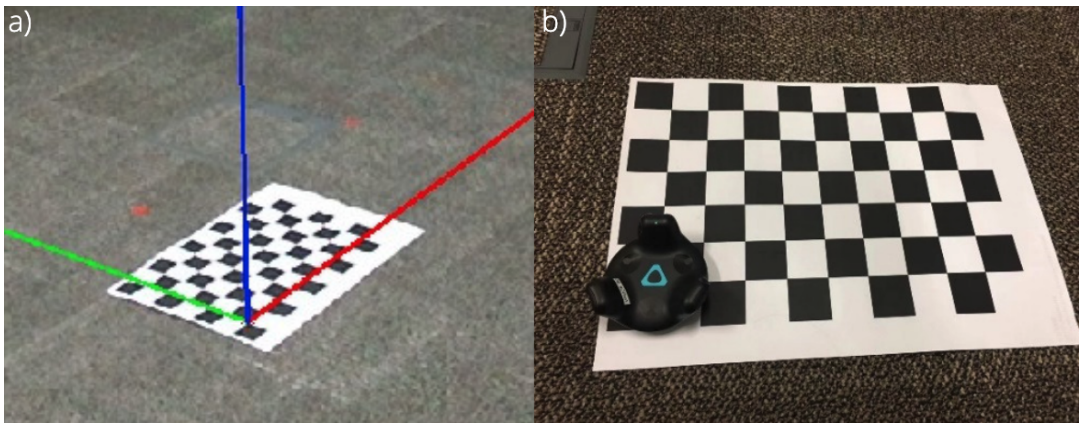


FIGURE 4.2: Calibration for Kinects and Vive system. a) Kinect and checkerboard, b) Vive tracker and checkerboard

.

Checkerboard to Vive system

After the relationship between the local coordinate system of each Kinect to the world coordinate system was required, a Vive tracker was used as a link between the Kinects and the Vive system. The local Y-axis of the tracker (as given by Unity) aligns with the index light direction. Since the Vive tracker is located on a corner of the same checkerboard (see Figure 4.2b), the checkerboard's coordinate system was manually configured with the Vive tracker's local coordinate system (with 10mm offset in the Z-axis direction).

Camera Offset Correction

Since there is some physical distance between the built-in RGB and depth cameras in Kinect, an additional calibration to avoid disparity error in the final results needs to be conducted before calibrating other devices. Otherwise, there is a clear and visible offset between the skeleton points (see Figure 4.3a). This offset can impact the final tracking results, such as introducing a positional shift of the body when the user rotates. Kwon et al. [69] provided a method to calibrate multiple Kinects with IR cameras, but it showed a limitation when calibrating heterogeneous tracking systems like the Vive and Kinect. To resolve the problem, the checkerboard-based calibration method was kept for the different types of coordinate systems. Figure 4.3b shows the results obtained after calibration.

To eliminate possible effects from other offsets, an additional calibration procedure was implemented. The participant was asked to stand in the center of the tracking area while facing Kinect 1. Then calibration matrices for Kinects 2, 3, and 4 based on Kinect 1 were calculated using three vertical joint positions on the torso of the user. Afterward, the joints were well aligned (see Figure 4.3b). Once the skeleton data was calibrated, the transformation matrix between the fused skeleton and the Vive coordinate system was calculated in the same way using the checkerboard to make the head transform from multiple Kinects consistent with the HMD.

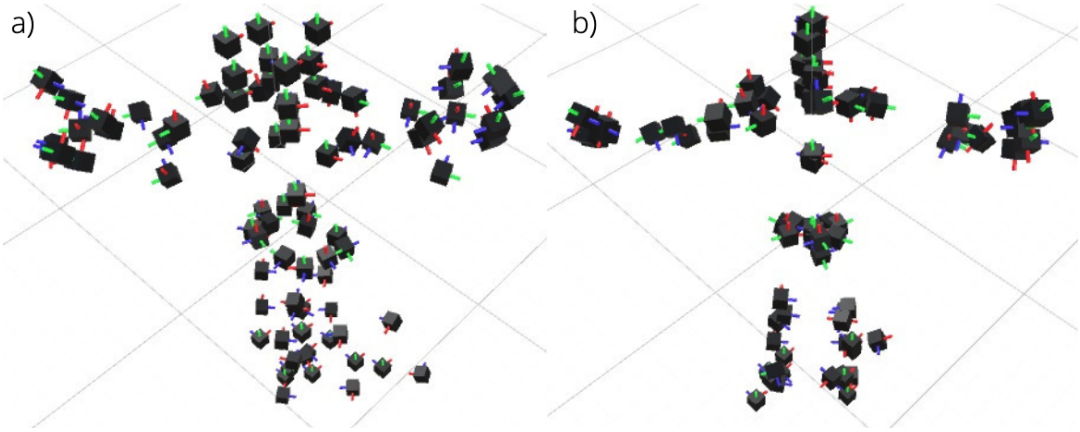


FIGURE 4.3: Additional calibration for the camera offset. a) Before additional calibration, b) After additional calibration. Notice the large number of cubes visible at the top (where the head would be) on the left, and the much smaller distance between them on the right, after additional calibration.

To stabilize the skeleton data of each Kinect, double exponential smoothing [60] was used recommended by Microsoft for jitter reduction and smoothing before sending the data to the server machine.

4.2.2 Refining Facing Direction

A stable facing direction is necessary to be able to provide a correct perspective for the user. 21 joint cubes were used to represent the user's skeleton and added a purple line that comes from the mid spine to indicate the facing direction (see Figure 4.4a). In Chapter 3, the facing direction was calculated when the tracking data needed to be processed every frame. The Left-Right

Swap (LRS) method was used, where a weighting factor was applied to each Kinect. The weighting factor calculation will be discussed later on data fusion. However, in Chapter 3, the facing direction sometimes suddenly flip in the opposite direction when the user rapidly rotated in the tracking area, as shown in Figure 4.4b. This swapping could affect the data fusion, as it determines when to invoke the LRS function, and controlling the avatar could be a problem due to the torso direction being pointed opposite to the head and limbs.

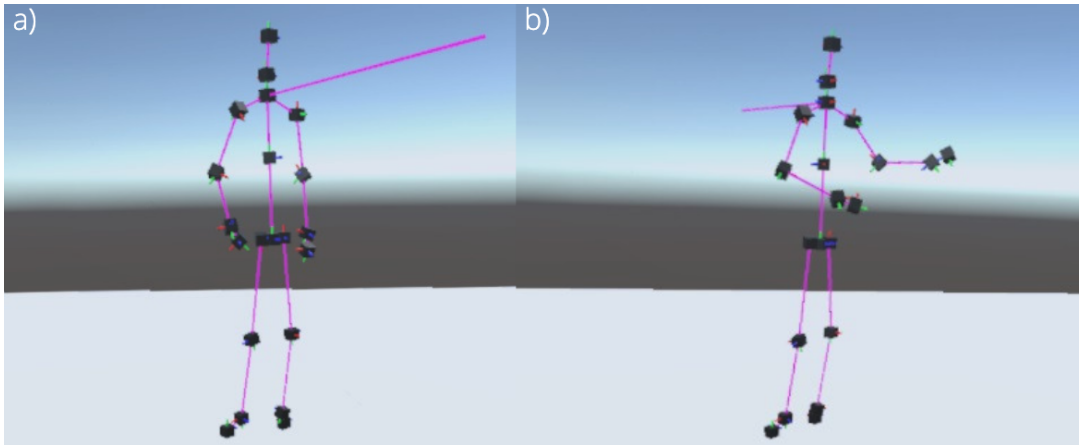


FIGURE 4.4: Two facing direction situations. a) Correct facing direction, b) Reversed facing direction

To address this problem, the direction of the HMD was used as a reference. As all the local coordinate systems from the heterogeneous devices (Kinect, Leap Motion, Vive) are calibrated using the same world coordinate system, the facing direction can be adjusted by combining the direction values from the HMD and multiple Kinects.

4.2.3 Data Fusion

The camera weighting approach plays an important role in data fusion, as it determines the data ratio that is coming from connected devices. The previous camera weighting method in Chapter 3 was used for the data fusion. However, It was found that jitter issues occasionally happened when the user lifted an arm while turning around, especially when the user faced specific directions. In order to identify the source of the jitter, several users were asked

to stand in the central tracking area and spin around. The skeleton from a third-person perspective was observed while the user saw if their action was synchronized or not from the first-person perspective. The results showed that the tracking quality was worse when the user was facing in a direction half-way between two Kinects. According to the previous data fusion method, the camera weighting for Kinect 1, 2, 3, and 4 was the same, although the tracking state for each joint was different. This meant that any bad tracking data would impact the overall fusion quality.

To explore the tracking quality from different angles, a test was implemented comparing a Kinect with three Vive trackers placed on the left shoulder, elbow, and wrist, respectively. As the Vive trackers and the Kinect were in the same coordinate system, the virtual representations of the Vive trackers were positioned close to a relevant joint of the fused Kinect skeleton. Four sets of data were collected when a user turned around in front of the Kinect with four angles 0, 90, 180, and 270 degree. This corresponds to the front view, the left (good side) view, the rear view, and the right (bad side) view of the user, respectively, with respect to the position of the Kinect. The users were asked to perform two movements, where the first movement involved the user lifting their left arm so that it was straight and aimed to the front while flapping vertically. The second movement involved the user lifting their left arm so that it was straight and aimed to the left while flapping vertically. The purpose was to observe the quality of the tracking state of different body parts under different occlusion conditions. Euclidean distance was used to calculate the difference between the relevant joints for data analysis of Vive trackers and Kinect, as detailed in Figure 4.5.

The Euclidean error was negligible when the user faced the Kinect (the front view) and was still small even when the user turned 90 degrees (the "good" side view) shown in Figure 4.5a. The worst condition was at 270 degrees (the "bad" side view), which fully occluded the left arms. Here, the Kinect could no longer recognize the front or backside of the user. Therefore, when the user turned 180 degrees (back view), the quality of the tracking was still good as the Kinect speculated the data was coming from the front side.

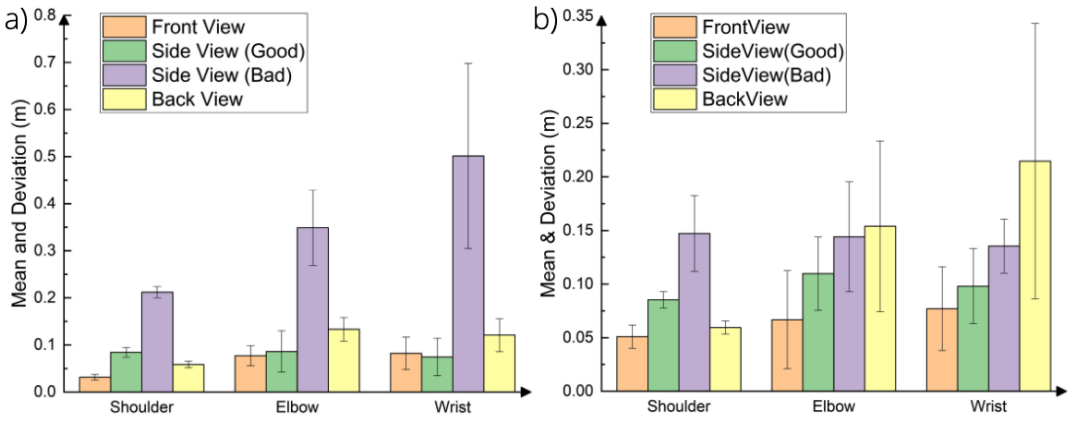


FIGURE 4.5: Tracking error comparison from one Kinect and Vive trackers. a) Arm lifted to the side, b) Arm lifted forward

The error in Figure 4.5b was different compared to Figure 4.5a, because the elbow and wrist were occluded by the shoulder joint when the user's back faced towards the Kinect. These results demonstrate that the central issue is occlusion, and any view where the tracking target, such as an arm, is occluded will likely be problematic and lead to unreliable tracking data.

Therefore, camera weighting for each Kinect needed to be reconsidered, especially when the user's good side was facing the Kinect. It is not reasonable to apply the same weighting to all the data from one Kinect as some body parts may be occluded, while others remain visible.

Weighting Factors for Improving Data Fusion

Two options for camera weight calculations are proposed for the "bad" tracking direction: 1) calculate camera weight distribution based on the different body parts, 2) reduce the weights from the back cameras to make the "good" tracking device (the front view) have a greater contribution. Two methods are presented below.

Method 1: Sub-region Weight Calculation (SWC)

The body of the user can be divided into three parts: left, right, and torso. The weights are different for arms and legs when the user is moving around in the tracking area, as shown in Figure 4.6.

- **Torso**: The torso is the middle section of the user, which includes the head, neck, the middle shoulder, the middle of the spine, and the spine base joint. The weighting factors in this section are decided by the previous method, which gives the highest weights to the front and back Kinects, and the lowest weights to the side Kinects.

- **Arms**: Other body parts can occlude the arms while the user freely moves around with a range of postures and gestures such as stretching, flapping, and bending the arms. Therefore, different situations (arms and legs occlusion situations as Figure 4.6) should be considered to assign the weight factors to joints such as the shoulders, elbows, and wrists.

- **Legs**: The weight calculations for the legs are done the same way as the arms, but using the facing direction instead of the body vector (a vector between two joints). The reason is that one leg is more likely to be occluded by the hips and the other leg when tracking using Kinects from the rear side. Therefore, it is reasonable to give higher weights to the front-side Kinects than the rear Kinects. The weighting curve for the arms and legs can be seen in Figure 4.8a

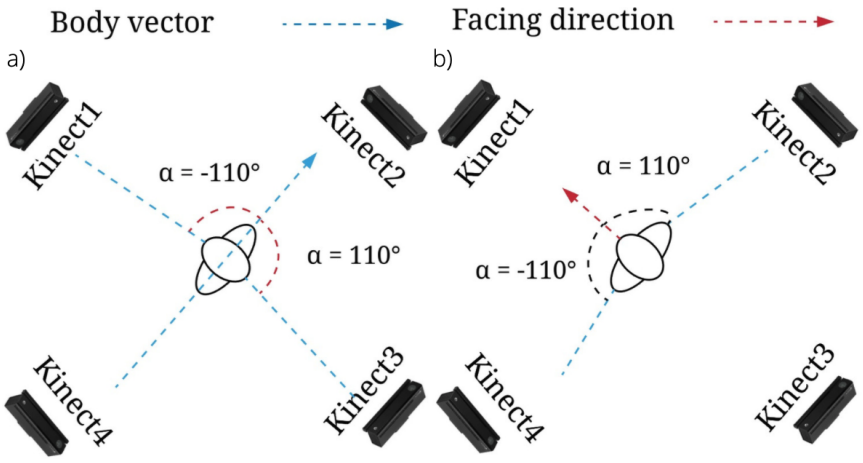


FIGURE 4.6: Method 1: Arm and leg weight calculations for each Kinect (right arm facing Kinect 2). a) Arm weight calculation, b) Leg weight calculation

Right arm was used as an example to test out the weight calculations. The right side of the user in Figure 4.6a is facing Kinect 2 and the data coming

from Kinect 4 should be allocated a lower weight for the right arm as the left body part occludes the right part. The camera weight from Kinects 1, 2, and 3 for right arm fusion should be the main data source as it is clearly visible from them. The weight calculation is defined by Equation (4.1), where λ represents the angle between the body vector (right and left shoulder pair) and the camera direction:

$$weighting = \begin{cases} 1, & -110° \leq \lambda \leq 110° \\ e^{\frac{100-|\lambda|}{40}}, & 110° \leq |\lambda| \leq 180° \end{cases} \tag{4.1}$$

Method 2: Improved Adaptive Weight Calculation (Improved AWC)
This method improved the weighting for each Kinect when the user is facing towards the established problematic direction, as shown in Figure 4.7a (half-way between two Kinects). According to the method in Chapter 3 (Method 3 in this Chapter), the camera weights were the same when the user faced the middle section, but the Kinects behind the user might not be in a good tracking state for the arms and legs. The new method linearly changes the weighting factors so that those from the backside tend to zero when the user turns 125 to 145 degrees. For example, the weighting for Kinect 3 and 4 decreases to 0 when the user is turned 35 to 55 degrees compared to Kinect 1, and the data fusion will be dominated by Kinects 1 and 2, as illustrated in Figure 4.7b.

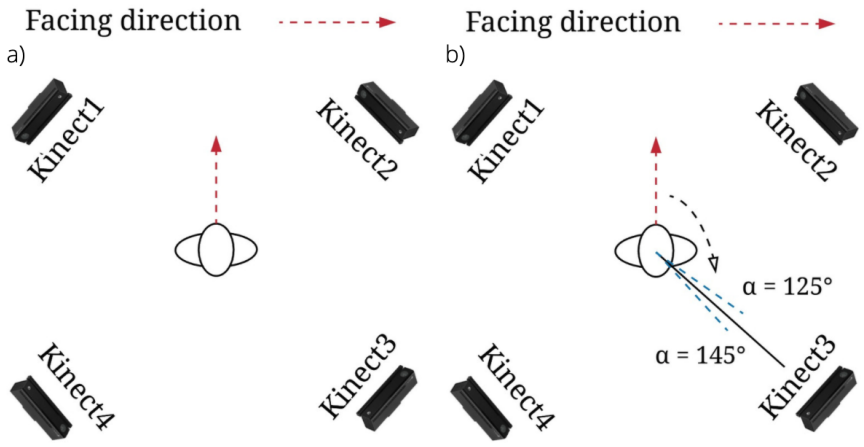


FIGURE 4.7: Method 2: Arm and leg weight calculations for each Kinect. a) The "bad" tracking direction, b) Camera weight calculation

The weight calculations were defined by Equation 4.1 according to the angle between the facing direction and the camera direction. The weight was normalized, and the resulting curves are shown in Figure 4.8b.

$$weighting = \begin{cases} \frac{|\lambda|-135}{45}, & 135° \leq |\lambda| \leq 180° \\ \frac{0.7\cdot(135-|\lambda|)}{18}, & 125° \leq |\lambda| \leq 135° \\ \frac{|\lambda|-90}{90}, & 90° \leq |\lambda| \leq 125° \\ \frac{90-|\lambda|}{90}, & |\lambda| \leq 90° \end{cases} \tag{4.2}$$

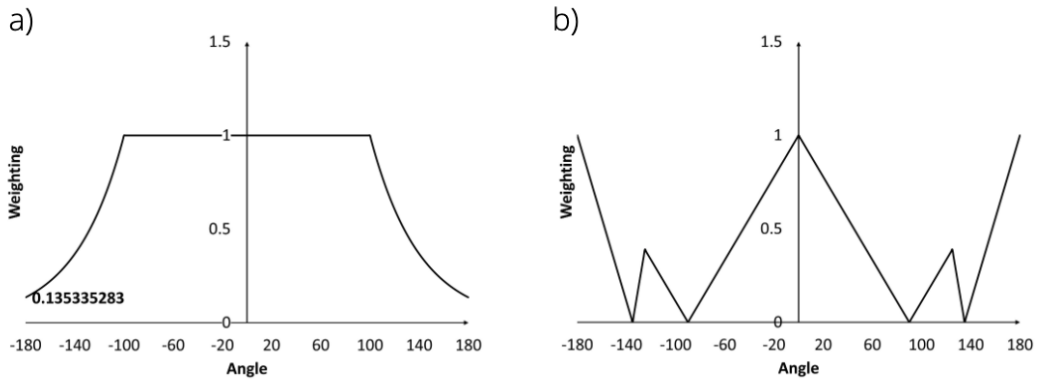


FIGURE 4.8: Weighting curve for each Kinect. a) Method 1: Sub-region weight calculation, b) Method 2: Improved adaptive weight calculation

Method 3: Adaptive weight calculation (AWC)

For the evaluation performed in section 4.3, the adaptive weight calculation (AWC) from the previous work in Chapter 3 will be used as a reference. This method is reprinted here in Equation 4.3:

$$weighting = \begin{cases} \frac{90-|\lambda|}{90}, & 0° \leq |\lambda| \leq 90° \\ \frac{|\lambda|-90}{90}, & 90° \leq |\lambda| \leq 180° \end{cases} \tag{4.3}$$

Position Data Fusion The position fusion method was verified in Chapter 3, which performed the best among three fusion algorithms. This method was adopted as defined by Equation 4.4.

$$\overline{P_j} = \frac{\sum_{i=1}^{4} \lambda_i^j \cdot w_i \cdot P_i^j}{\sum_{i=1}^{4} \lambda_i^j \cdot w_i}, j \in \{0, 1, ..., 24\} \tag{4.4}$$

where $\overline{P_j}$ is the fused position of the j^{th} joint, P_i^j and λ_i^j are the position and tracking state of the j^{th} joint from the i^{th} Kinect, and w_i is the weight for the i^{th} Kinect.

Rotation Data Fusion

The rotation data from each Kinect represents how the joints of the user rotate relative to the camera, which is important when controlling an avatar. In general, there are two ways to implement avatar control. One is based on positions to calculate the bone rotation between two joints. The other uses the rotation data from the Kinect device. Therefore, the rotation data from multiple Kinects still needs to be considered for avatar control. The fusion procedure is as follows:

(1) Calculate weights for each quaternion from the relevant Kinect

$w_1 = \lambda_1 \cdot cw_1$, $w_2 = \lambda_2 \cdot cw_2$, $w_3 = \lambda_3 \cdot cw_3$, $w_4 = \lambda_4 \cdot cw_4$;

λ_i is the joint tracking state and cw_i is camera weight from each Kinect.

(2) Calculate interpolation terms from the weights

$i_1 = \frac{w_1}{w_1+w_2}$, $i_2 = \frac{w_3}{w_1+w_2+w_3}$, $i_3 = \frac{w_4}{w_1+w_2+w_3+w_4}$;

(3) Quaternion fusion

if $q_1 \cdot q_2 < 0$ (The rotation of q_1 and q_2 are opposite) **then**

$q_1 \leftarrow -q_1$;

$q_{1-2} \leftarrow slerp(q_2, q_1, i_1)$;

else

$q_{1-2} \leftarrow slerp(q_2, q_1, i_1)$;

end if

if $q_1 \cdot q_3 < 0$ **then**

$q_{1-2} \leftarrow -q_{1-2}$;

$q_{2-3} \leftarrow slerp(q_{1-2}, q_3, i_2)$;

else

$q_{2-3} \leftarrow slerp(q_{1-2}, q_3, i_2)$;

end if

if $q_{2-3} \cdot q_4 < 0$ **then**

$q_{2-3} \leftarrow -q_{2-3}$;

$q_{final} \leftarrow slerp(q_{2-3}, q_4, i_3)$;

else

$q_{final} \leftarrow slerp(q_{2-3}, q_4, i_3)$;

end if

q_i is the quaternion from i^{th} Kinect and q_{i-j} is the quaternion from the i^{th} Kinect to the j^{th} Kinect.

4.3 Evaluation

In this section, an evaluation of the tracking quality and the errors (measured by the standard deviation) of the retrieved data from multiple tracking devices for avatar skeleton rigging in VR are presented. The evaluation targets the following three parts: facing-direction adjustment, camera-weighting method, and rotation-fusion procedure.

4.3.1 Facing-direction Adjustment

To verify the success of the facing direction stabilization, the proposed method was compared to the prior method in Chapter 3, where the participants were asked to rotate rapidly. Participant's facing direction were represented using two vectors that come from the Kinect fused data (red) and from the HMD (blue) (see Figure 4.9). The user stood in the tracking area, and the fused

skeleton for the user was visualized in VR. The facing direction from the fused system and the HMD were recorded when the user turned around until the facing direction line changed to the reverse direction (see Figure 4.4b).

The trajectory of the two vectors was visualized using a polar diagram, similar to a top-down view (Figure 4.9a). In the previous method, the two vectors were almost overlaid before the value changed at the 10-second mark, after which the signals no longer matched. The difference between these two vectors was calculated and is presented in Figure 4.9b. The graph shows that the difference curve (grey) stayed around 0 degrees before the 10-second mark. The value then changed to 180 degrees after 10 seconds, which proves that the facing direction swapped.

The HMD variation curve in Figure 4.9b was smoother than the data from the Kinect, which means that more samples were taken at the same time by the Vive system compared to the Kinect. Therefore, it is more reliable to adjust the facing direction by using the HMD when the direction abruptly reverses.

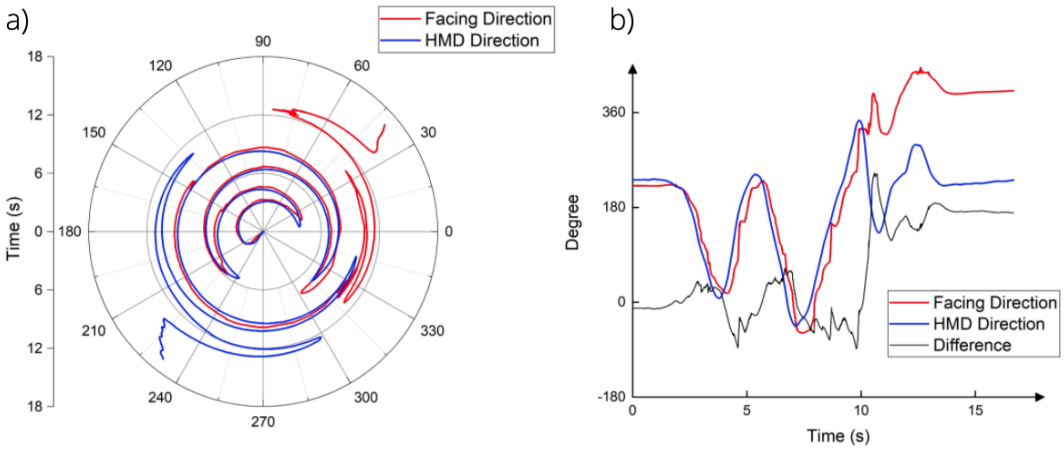


FIGURE 4.9: Comparison between the two facing directions during a fast rotation (using the previous method from [138]. a) The polar diagram of the facing direction, b) The difference between the calculated facing direction and the HMD direction

In the second test, the updated method was used to calculate the facing direction. The data acquired was recorded using the same method as with the previous approach. In Figure 4.10a, the facing direction reversed twice during the recording, and the value was corrected by the HMD direction immediately, which can be seen from the trajectory. The timestamps in Figure 4.10b show

the abrupt changes, and the difference between the two vectors is -90 to 90. The user may have turned their head while rotating during the process, which can cause the vector difference. Therefore, it is concluded that the new facing direction calculations are significantly more stable than the previous approach.

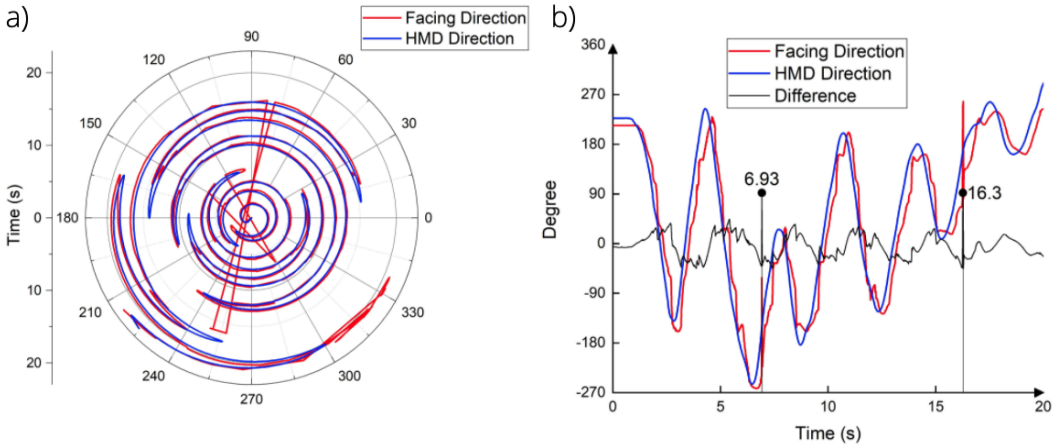


FIGURE 4.10: Comparison between the two facing directions during a fast rotation(using my proposed method). a) The polar diagram of the facing direction, b) The difference between the calculated facing direction and the HMD direction

4.3.2 Camera-weight Comparison

The forearm of a user may sometimes be incorrectly lifted when the user lifts their other arm and faces in a specific direction with the current camera-weighting method. To test how well the new camera-weighting methods can solve these issues, the data was collected from eight angles using the three camera-weighting distribution methods: (1) SWC, (2) Improved AWC, and (3) AWC.

Two experimental comparisons were made: (1) lift the right arm and measure the data from the left arm while stationary as in Figure 4.11, and (2) lift the right arm and flap it vertically. Data were collected both from three Vive trackers (used as a reference) placed on the left shoulder, the elbow, and the wrist, along with the fused data from the three camera-weighting methods.

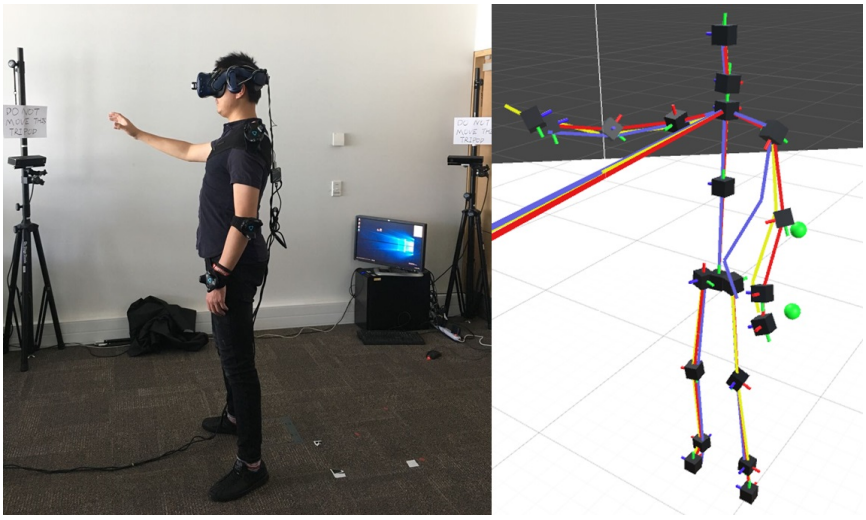


FIGURE 4.11: Three skeletons with the three camera weighting methods for an arm lift

In Figure 4.11, we can see three skeletons, one for each camera-weighting method, where the colors represent BLUE for AWC, YELLOW for SWC and RED for the Improved AWC. In the skeleton image in Figure 4.11, the left arm of the user did not coincide with the green spheres which represent the data from the Vive trackers placed on the left arm. From Figures 4.12a and 4.12b, it can be seen in the Improved AWC case that the error between the left wrist and the tracker was low in directions 1, 3, 5 and 7. This is because the user was directly facing one of the Kinects. However, in directions 2, 4, 6, and 8, as the user was facing half-way between two Kinects, the errors were much higher.

AWC had, by far, the largest error in these problematic directions. This supports the claim that, given the equal weighting applied to all Kinect data, the poor data from one Kinect will negatively impact the data fusion result. The SWC avoided the problem of using a single weighting factor for all data from one Kinect by, for example, weighting higher the left body part data coming from the Kinects on the left. This decreased the error when compared to AWC, but it still was not sufficient. Looking at Figure 4.7a for instance, when the user faced half-way between Kinects 1 and 2, the left arm fusion data mainly came from Kinect 2 and 3, but Kinect 4 may mistakenly recognize the lifted right arm as the left arm. It would then contribute a high weighting for

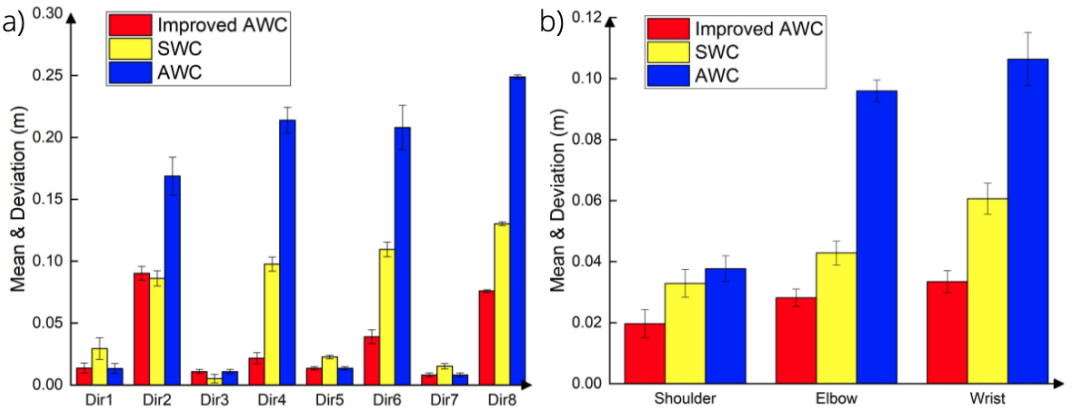


FIGURE 4.12: Error comparison between Vive trackers and the three camera weighting methods. a) The mean error between the left wrist and the Vive tracker, b) The mean error between the left arm and the Vive trackers

the left forearm, which was the reason why the virtual left arm was lifted when the user lifted the right arm. The Improved AWC approach performed the best with the lowest error, demonstrating that the special weighting considerations were successful. The camera weights from the backside Kinects decreased to 0 linearly when the user faced half-way between the two front Kinects. Therefore, the risk of bad tracking from the backside Kinects was minimized.

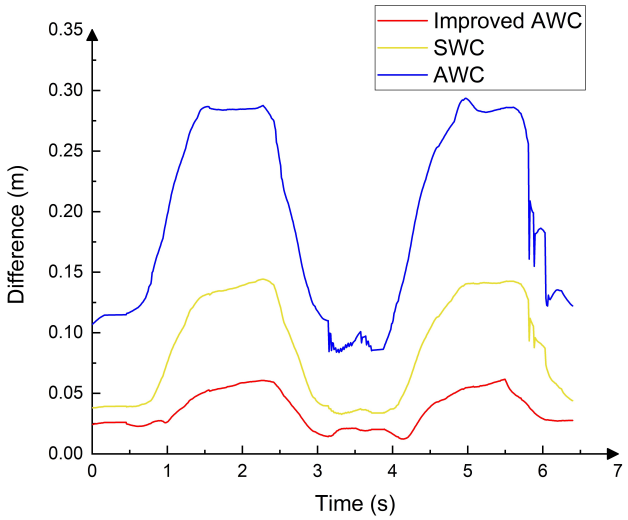


FIGURE 4.13: The difference between the three weighting methods and the Vive tracker for the wrist for experiment 2

Figure 4.13 describes the difference between the wrist and the relevant Vive tracker during continuous arm flapping when the user faced in a problematic direction. The wrist performed the worst as shown in Figures 4.11 and 4.12.

From the analysis above, it is concluded that the Improved AWC method increased the data fusion quality, and should, therefore, be preferred.

4.3.3 Rotation-data Fusion

To test the accuracy of the rotation-data fusion, the Vive tracker was attached on the left elbow of the user and tested the rotational difference between two relevant joints. To make sure the comparison was between the same axis, the user was asked to lift and stretch the arm to check the relevant pointing direction. This shows that the Y-axis from the Kinect elbow and the Z-axis from the Vive tracker are aligned with each other and point in the same direction relative to the world coordinate system. The user was asked to twist the arm, and two states (before and after the twist) were recorded in Figure 4.14.

It can be seen from Figure 4.14a that the rotation data from the Kinect and the Vive tracker are centered around 60 and 65 degrees, respectively. The difference and mean can be seen in Figure 4.14b, which shows that the fused rotation data had a difference of roughly 5 degrees on average.

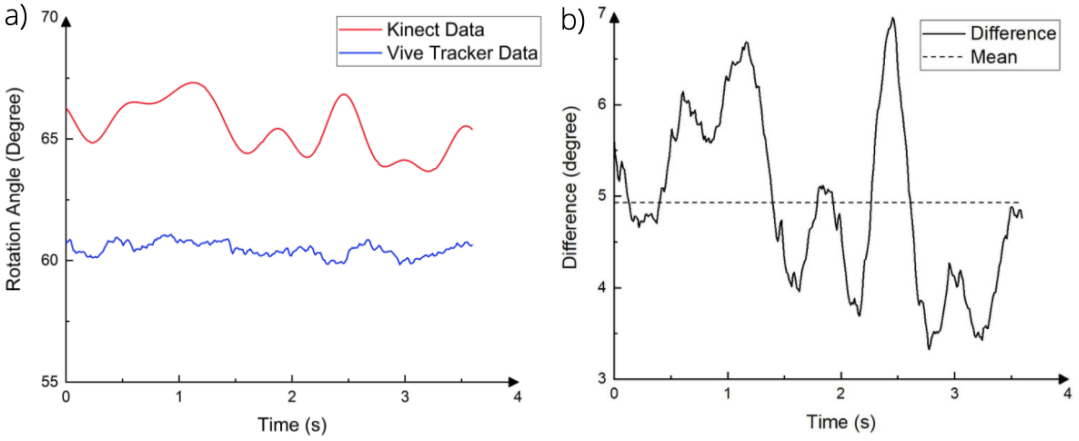


FIGURE 4.14: A comparison between the fused Kinect joint and the Vive tracker for rotation comparison. a) The rotation angle between the two raw data streams, b) The rotation difference and mean

4.4 Leap Motion and Fused Kinect Integration

For providing an articulated tracking system including finger movement, a Leap Motion hand tracker was attached to the front of the HMD. The data was fused from multiple Kinects. Hand tracking from a Kinect is not sufficient, as it jitters significantly, even when the user faces the camera. Though the thumb, hand tip, and hand open state data can be retrieved from the Kinect, this is not fine-grained enough to be applied to an avatar or used for interaction in VR. Instead of using hand and wrist data from the Kinect, Leap Motion data can be integrated with the fused skeleton to support rich hand gestures and finger movement. The Leap Motion coordinate system was aligned with the Vive coordinate system as described above. Once the Leap Motion was calibrated with the HMD and multiple Kinects, all the data are fused and visualized in Unity as shown in Figure 4.15a.

Although the hand tracking by Leap Motion is better than that from the Kinect, the tracking range and angle are quite small (60cm above the controller, by 60cm wide on each side, by 60cm deep[5]) due to hardware restrictions. Therefore, the user must put their hands in front of the HMD to position them in the tracking area. To make it work well in the fused system, the data was automatically substituted within the fused skeleton system. The hand, finger, and wrist data come from the Leap Motion when the hands are in its tracking area. Otherwise, the data is used from the fused Kinects, as shown in Figure 4.15b.

4.5 Conclusions and Future Work

In this chapter, a set-up using multiple Kinects is introduced for robust and accurate full-body 3D skeleton tracking together with Leap Motion integration into a Vive system for VR. A calibration method was suggested to synchronize heterogeneous devices easily using a traditional checkerboard marker. New camera weighting methods were proposed and compared with the previous

[5]http://blog.leapmotion.com/hardware-to-software-how-does-the-leap-motion-controller-work/

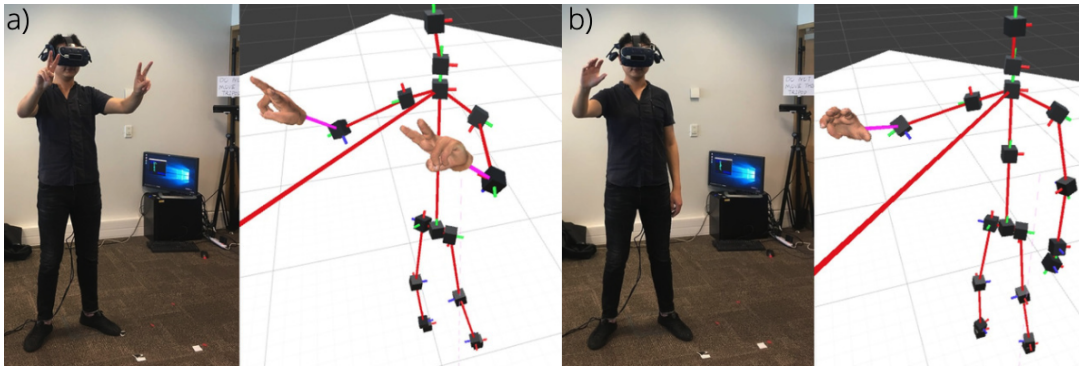


FIGURE 4.15: Integration of the fused Kinects and the Leap Motion. a) Full body with hand tracking from the fused Kinects and Leap Motion, b) Different data sources for left and right hand tracking

approaches. The results showed that the improved AWC method proposed in this chapter could tackle several tracking issues. The results of the rotation fusion tests show that the system has good accuracy compared to the Vive tracker. The Leap Motion data was integrated with the fused skeleton system, supporting hand gestures in VR interaction scenarios. In the next chapter, I will introduce the avatar control system and answer the question: How and to what extent can the integration of multiple depth sensors of the avatar control system support better communication?

Chapter 5

The Effect of Avatar Expressiveness on Communication in VR

Fully-tracked avatars with rich hand gestures in VR are required for good communication, especially in social scenarios. In this chapter, I focus on increasing the behavioral fidelity of a participant's virtual body representation. To investigate the impact of a full-body avatar control system with hand gestures, I compared it against a controller-based avatar system (partial-body tracking with limited hand gestures). A VR interview simulation was designed for a single user to measure the effects on presence, virtual body ownership, workload, usability, and perceived self-performance. Specifically, the interview process was recorded in VR, together with all the verbal and non-verbal cues. Subjects then took a third-person view to evaluate their previous performance. The results show that the full-articulated avatar control system increased virtual body ownership and also improved the user experience. Besides, users rated their non-verbal behavior performance higher in the full-body avatar system. This work was presented as a full paper [139] in the 25th ACM Symposium on Virtual Reality Software and Technology (VRST 2019), which was held in Sydney, Australia, from November 12-15, 2019.

5.1 Introduction

At present, there are many different solutions for providing embodied virtual experiences. The avatar control mode may impact the user's performance,

body ownership, or social behavior in multi-user social scenarios, which led me to explore how different methods for controlling an embodied avatar to support communication.

To understand whether different levels of articulation of a virtual avatar impact communication behavior, the avatar control system described in Chapter 4 was used to provide full-body and hand tracking. To minimize the number of wearable devices while providing sufficient tracking quality, the data was fused from multiple commercial tracking sensors from four opposite directions around the user. the data was used to control a virtual avatar, removing the need to wear any sensors. To explore the effects of different avatar control strategies in terms of presence, virtual body ownership, workload, usability, and communication performance in VR, a virtual interview experiment was implemented between depth-sensor-based and a controller-based avatar control approaches.

The the subjects used both systems in a mock interview process. To assess and improve the self-evaluation experience, I went through a review session, and the user could review their previous performance in VR from a third-person point of view. The remainder of the chapter describes the approach in detail.

5.2 Methods

An avatar control system was built based on the work in chapter 4. The user can move freely with full-body (21 joints, including the torso, arms, and legs) and hand gesture tracking (19 joints with pointing, grasp, and pinch). A study was designed to evaluate the effects on communication behavior between the two experimental conditions: (1) A motion-capture tracked avatar, providing an embodied representation of the user with full-body and hand-gesture tracking, and (2) Controller-based avatar control system, using the tracked HMD and controllers to get an embodied representation with upper-body tracking and a single pointing gesture. Both conditions included a

shared workplace with a virtual interviewer. More details can be found in Section 5.2.3.

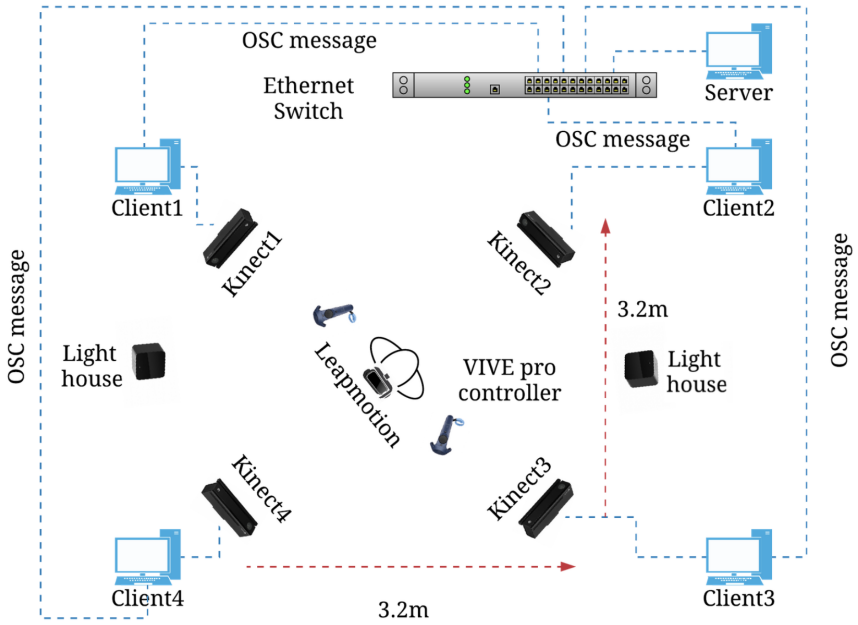


FIGURE 5.1: Setup for the depth-sensor-based avatar control system. Four Kinects tracked the user's body and Leap Motion tracked hands in real-time.

5.2.1 System Overview

Hardware Overview

The tracking system described in Chapter 4 was installed around a 3.2m x 3.2m square tracking area. The Kinects were adjusted to 1.2m above the floor to avoid infrared radiation interference between the Kinects and the HMD. Each Kinect was driven by a client machine, which was an Intel NUC (Intel Core i5-8259U at 2.3 GHz, 8GB RAM, and Iris* Plus Graphics 655). An HTC Vive Pro with two second-generation Lighthouse sensors was driven by a server machine, which was a Windows 10 desktop computer (Intel Core i7-7700K at 4.2 GHz, 32GB RAM, and NVIDIA GeForce 1080Ti). A Leap Motion sensor was attached to the front of the HMD with a USB cable to the server. All four client machines and the server machine were connected to a Gigabit Switch (NETGEAR GS110MX) through Ethernet cables for network data transmission.

Software Overview

Development Tools: The software was developed using the Unity game engine version 2017.1.1f1 [125] with SteamVR for Unity [25], and Leap Motion plugin for Unity (Core 4.4.0) [74]. The skeleton data from each Kinect was wrapped as an Open Sound Control (OSC) message and transmitted through the network. The UniOSC plugin [130] was used to handle the OSC messages received on the server machine.

Virtual Avatar Representation: Since the sense of ownership can be induced even using a body part that is not your own [17] and my goal is not related to the personalized appearance of the avatar, three generic avatar models were created using MakeHuman software [76] in this study. Two avatars were designed for participants with standard body size, and different heights [22] for female (1.65m) and male (1.77m) characters. As the user study is a mock virtual interview, the virtual interviewer was customized with a similar appearance and body size as the lab lead. To avoid any expected confounds from facial expressions since I used an interview scenario, plausible expressions were developed, including mouth movement, eye movement, and blinking for both conditions. Blender 2.79b [14] was used to add blend shapes on the avatars and used the SALSA plugin for Unity [122] to customize three sets of blend shapes to represent the open mouths as small, medium, or large in shape, which were triggered by the loudness of the microphone input. The eye-gaze direction was the same as the head orientation of the HMD, and the virtual avatar performed random eye blinking. The skeleton was droven in Unity using all these procedures, including the tracking system. The participants had different heights and body sizes. An additional calibration step was carried out to apply the predefined character to each participant. The participants were asked to stand still and make a "T-pose" before the experiment. Their height and arm lengths were measured and used to auto-scale the size of the virtual character

Bandwidth and latency

Four client machines continuously streamed the serialized body-frame data to the server machine at a rate of 1.5Mbps over Ethernet. There were three sources of system latency: (1) Data pre-processing on the client machine. It took less than 1ms for the client machine to serialize the body-frame data into an OSC message before sending it to the network once the Kinect detected the user in the tracking area. (2) Data transmission in the network. It took less than 1ms for message transmission. (3) OSC message handling and avatar control rendering in the Unity game engine. The UniOSC plugin was used to process the OSC messages received on the server machine and to deserialize the data for the fusion. As this plugin relies on the game engine, it took up to 30ms from data receipt and fusion to avatar rendering. Therefore, the latency of the system is less than 32ms in the worst case.

5.2.2 Participants

25 participants (13 male, 12 female) were recruited from University of Canterbury through advertisements posted on campus and University social media platforms. They were aged 18-35 ($M = 26.2$, $SD = 4.5$). Participants were asked about their familiarity with VR using a 5-point Likert scale, from 1 (never), 3 (a few times a month), to 5 (daily use). The participants generally had moderate experience for using VR ($M = 2.56$, $SD = 1.19$). The frequency of Social VR platform use was never (72%), a few times a year (20%), a few times a week (4%), and daily (4%). From the demographic information, most participants had VR experience, but only 28% of subjects had tried social VR applications before.

5.2.3 Study Design

A 1x2 within-subjects design was used with an interview scenario in a virtual office. For each experiment condition, the participants experienced two sessions: an interview session and a review session. In the interview session,

two specific tasks were provided. After the interview session with a virtual interviewer, the participant watched their recorded interview from the third person perspective for self-evaluation purposes in the second (review) session. The condition order was randomized using Research Randomizer [131] to avoid ordering effects.

Conditions

The Controller-based Avatar Control System (CB-ACS) - In this condition, the virtual character was driven by tracking the HMD and two controllers. The Final IK for Unity [98] plugin was used to calculate and estimate the positions and rotations of the joints of the body, excluding the head, and left and right hands. In this case, the virtual hands and arms moved when the participant moved the controllers. The two legs of the virtual character moved automatically when the translation of the HMD changed, and the step width was adjustable. As there was no finger tracking in this condition, a pointing gesture was made when the participant pulled the trigger button.

The Depth-sensor-based Avatar Control System (DSB-ACS) - In this condition, the motion tracking data came from the Kinect+Leap Motion system described above (Chapter 4). All tracked joint data was fed into Unity for avatar control. The Leap Motion has a limited field of view (60cm vertical x 60cm horizontal x 60cm deep) [23]. Therefore, the data for the elbows, wrists, and hands switched to the Kinect sensors whenever the hands were outside of the Leap Motion tracking area.

Two sessions in this experiment

First session: Be an interviewee - The effect of the avatar control approach on the user's behavior can depend upon the interview questions and tasks that the participant needs to accomplish. All the interview questions and the two tasks were carefully designed to make sure that the participant could employ the body postures and hand gestures naturally and intuitively.

Task 1: Answering interview questions The participant needed to answer a set of interview questions from a virtual interviewer. The tested and used interview questions by Villani et al. [132, 133] were adopted. For each condition, three questions were adopted from the question sets. The time for each answer was two minutes. A "stop" animation and relevant audio were used to remind the user to stop and answer the next question. Two interview question sets were prepared, and one was chosen at random for a given participant:

- **Set 1:** What is your greatest weakness? Where do you see yourself in five years? Tell me about a time when you used your skills of persuasion to convince someone of your ideas.
- **Set 2:** How will your greatest strength help you perform? What are your expectations and goals? Let us talk about your personality. What are three adjectives that best describe you?

Task 2: Route-planning Task After the interview questions, the participant was asked to complete a route-planning task by giving directions to the interviewer while referring to a virtual map present in VR. The map was shown on a "TV screen" placed on a nearby cabinet. The participant had to take a few steps to get close to the TV screen. The height of the TV top to the floor was around 1.7m. In the task, the participant had to describe a path from a given starting point (red circle) to an endpoint (blue circle) as Figure 5.2. I expected participants would use many social cues, such as finger or hand gestures, and body postures to confirm things with the virtual interviewer.

The non-verbal behavior of participants, such as body movements, hand gestures, mouth movements, and eye blinking data, was recorded at ten frames per second. This parameter could be set at a higher level, but it consumes CPU and RAM resources, which can slow the system when storing frame data of the avatar animations. The verbal-behavior audio was recorded at a 44.1kHz sample rate to guarantee high quality in the review session. All the interview questions and instruction audio was recorded in advance, as well as relevant animations by a native speaker. In the interview session,

FIGURE 5.2: The maps in the route planning task

the participant faced the virtual interviewer, and related animations were triggered manually by a researcher pressing specific keys on the keyboard. The session for the whole process was recorded automatically in Unity. Then the recorded timeline was used for playback in the review session.

Second session: Review interview from third-person perspective - In this session, the participants watched their interview process through an HMD. They could walk and turn around in this session to review the interview from a third-person view. The whole procedure was replayed automatically according to the recorded timeline. The participants were asked to focus on the verbal and nonverbal behavior of themselves for the questionnaire administered afterwards.

Hypotheses

I expected better avatar control and reduced encumbrance for holding the controller from the system. Based on these expectation, and previous related work in the field, several hypotheses were formulated.

- H_1: Depth-sensor-based avatar control will provide participants with a deeper sense of presence compared to controller-based avatar control.
- H_2: Participants will feel lower mental workload when using depth-sensor-based avatar control compared to controller-based avatar control.

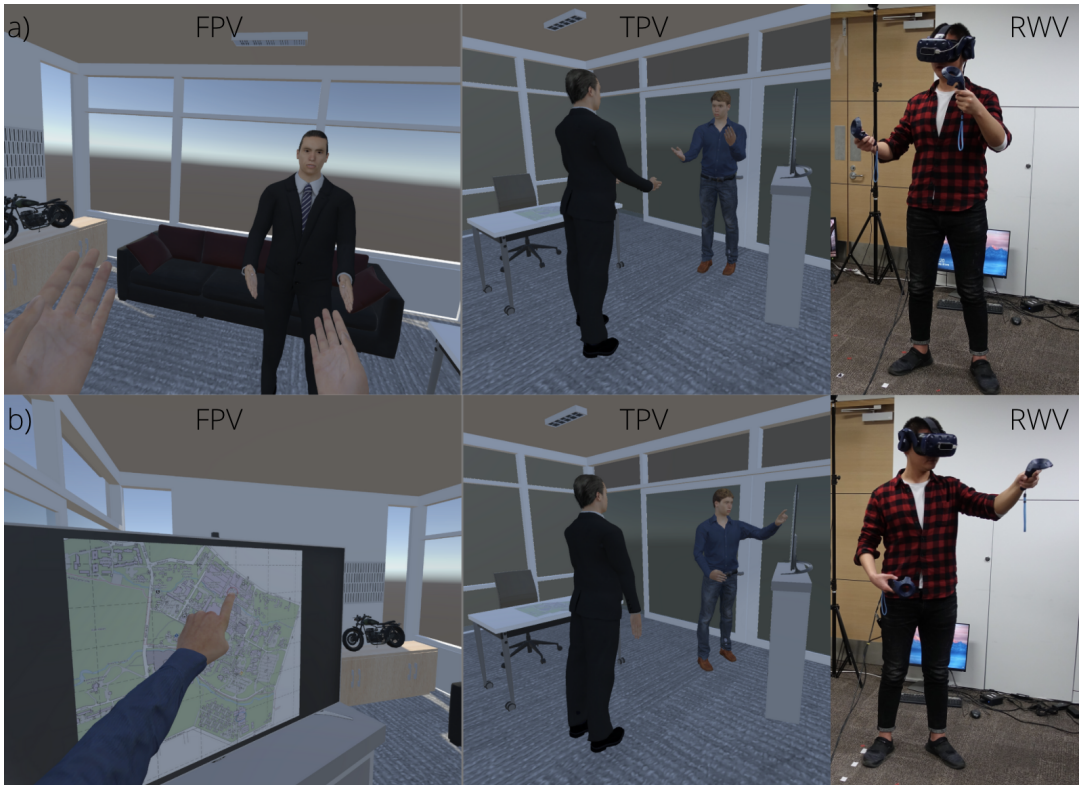


FIGURE 5.3: The virtual interview experiment in Controller-based avatar control condition with first-person view (FPV), third-person view (TPV), and real-world view (RWV): a) Task 1: Answer the questions, b) Task 2: Route planning task

- H_3: Using the depth-sensor-based avatar control system will result in a higher sense of body ownership and agency compared to controller-based avatar control.
- H_4: Depth-sensor-based avatar control will provide participants with a better user experience during virtual social communication compared to controller-based avatar control.
- H_5: Participants will have better self-rated performance in terms of communication behavior by using depth-sensor-based avatar control compared to controller-based avatar control.
- H_6: Participants will prefer to use depth-sensor-based avatar control over controller-based avatar control.

FIGURE 5.4: The virtual interview experiment in Depth-sensor-based avatar control condition with first-person view (FPV), third-person view (TPV), and real-world view (RWV): a) Task 1: Answer the questions, b) Task 2: Route planning task

5.2.4 Measures

Data were collected in two ways. Most subjective questionnaires were filled out after the interview session, which used a first-person perspective. As the participant could move around during the review session in the virtual environment from a third-person point of view, the self-evaluation questionnaire was filled out after the review session.

First-person Perspective

Dependent variables such as the sense of presence were assessed using the Igroup Presence Questionnaire (IPQ) [109], and the sense of body ownership and agency were measured using a questionnaire about avatar embodiment [47]. As there is a route planning task, the workload was assessed using NASA TLX [50]. The System Usability Scale (SUS) [19] was adapted to compare the usability of the two avatar control methods.

Third-person Perspective

To verify H_5, participants answered the following custom questions and scored themselves (0-100) in terms of verbal and non-verbal behavior after the review session.

"Think about what you saw when you watched the replay of your interview."

- *(1) How realistic was your non-verbal behavior: body posture and hand gestures?*
- *(2) How realistic was your verbal behavior?*

User Preference

Finally, a set of post-experiment questions were created for comparison between depth-sensor-based and controller-based avatar control methods in terms of ease of use and preference.

5.2.5 Procedure

After the introduction of the experiment, the participant signed the consent form and filled out the demographic survey on a laptop. The researcher explained how to use the devices involved in the study, helped the participant put on the HMD, and asked them to familiarize themselves with the controllers or hand tracking devices. The participants were asked to walk around to practice how to control the virtual avatar in the two conditions. They then spent one minute looking around the virtual environment to familiarize themselves with the furniture and layout. This was done to reduce the risk of distraction during the experimental tasks.

Participants were then positioned in the center of the tracking area and asked to start Condition one, Session one from the first-person point of view. The whole process was recorded and stored as animations. The audio from the participant was also recorded from the built-in microphone of Vive Pro. After the first session, the participant was asked to fill out several surveys on the laptop. In the next session (Condition one, Session two), the virtual camera in the scene changed to a position near the virtual interviewer. The

animations and audio were loaded for replay, and the participant watched the previous interview in VR from a third-person point of view. A self-evaluation survey was filled out by the participant after the replay session. The process was repeated for the second condition.

After the second condition, participants were given one additional survey to gather information about their preference and ease of use of the avatar control schemes. The researcher then performed an experimental debrief with the participant and encouraged them to write comments about the two systems, discuss their survey answers, and talk about their general impressions of the two conditions.

5.3 Results

In this section, the results of the effects of the depth-sensor-based and controller-based avatar control approaches are provided. For the analysis, 25 participant data sets were used. As I described in the user study section, the study was implemented as a 1x2 within-subjects design. Paired samples t-test was used to analyze subjective measures for presence, virtual body ownership illusion (VBOI), agency, workload, usability, and self-evaluation. For significance testing, a confidence value of $\alpha = 0.05$ was used.

5.3.1 First-person Perspective

Presence and Workload

Presence was measured from the IPQ with four components: General Presence (GP) ($t(24) = 0.558, p = 0.582$), Spatial Presence (SP) ($t(24) = 1.785, p = 0.087$), Involvement (INV) ($t(24) = 1.894, p = 0.07$), and Realism (REAL) ($t(24) = 1.272, p = 0.215$), and the overall workload ($t(24) = 1.361, p = 0.186$) was measured from NASA TLX. From the results presented in Table 5.1 and Figure 5.5a, 5.6a, we can see that there were no significant differences (p>0.05) between the depth-sensor-based and controller-based avatar control approaches.

TABLE 5.1: Statistical results for Presence and Workload

	Presence				**Workload**
	GP	SP	INV	REAL	
t-test	p=0.582	p=0.087	p=0.070	p=0.215	p=0.186
CB-ACS (M, SD)	(4.4, 0.9)	(4.2, 1.0)	(3.7, 1.1)	(2.9, 1.0)	(42.7, 13.5)
DSB-ACS (M, SD)	(4.5, 1.1)	(4.5, 0.7)	(4.1, 1.0)	(3.2, 0.7)	(46.0, 12.7)

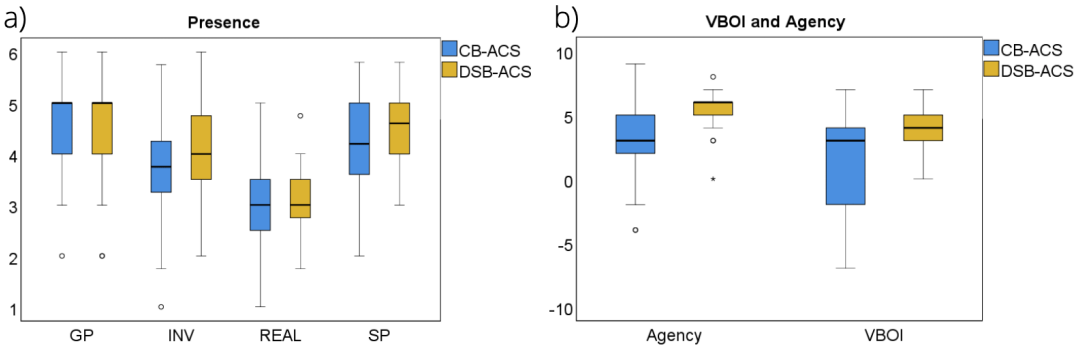


FIGURE 5.5: Presence, VBOI, and Agency

Virtual Body Ownership Illusion, Agency, and System Usability

The VBOI and sense of agency were measured from the avatar embodiment questionnaire. As there was no mirror placed in the virtual environment, Q4 ("I felt as if the virtual__I saw when looking in the mirror was my own__") and Q5 ("I felt as if the virtual__I saw when looking at myself in the mirror was another person") were removed. System usability was measured using the SUS. From Table 5.2 and Figure 5.5b, it is clear that there was a significant difference (p<0.05) between the depth-sensor-based and controller-based avatar control approaches in terms of VBOI ($t(24) = 3.385, p = 0.002$), agency ($t(24) = 3.748, p = 0.0009$), and usability ($t(24) = 3.313, p = 0.0029$).

TABLE 5.2: Statistical results for VBOI, Agency, Usability, and Performance. Bold indicates statistical significance.

	VBOI	**Agency**	**Usability**	**Performance**	
				NVC	VC
t-test	**p=0.002**	**p=0.0009**	**p=0.0029**	**p=0.008**	p=0.37
CB-ACS (M, SD)	(1.2, 4.0)	(3.0, 3.2)	(64.1, 16.2)	(52.1, 23.7)	(68.6, 23.6)
DSB-ACS (M, SD)	(3.8, 2.1)	(5.4, 1.6)	(74.0, 10.6)	(66.9, 14.5)	(72, 18.6)

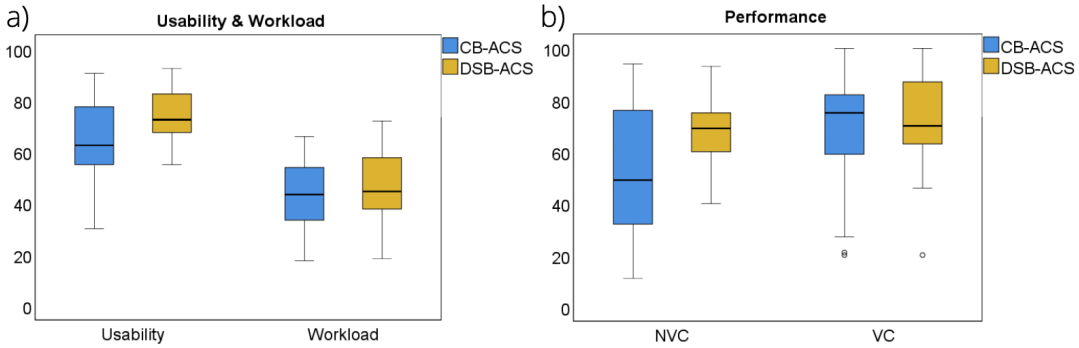


FIGURE 5.6: Usability, Workload and performance

5.3.2 Third-person Perspective

Self-evaluation

From the results in Table 5.2 and Figure 5.6, we observe an interesting outcome. There was a significant difference (p<0.05) between the two avatar control approaches in the interview review session in terms of non-verbal behavior ($t(24) = 2.878, p = 0.008$). This indicates that participants preferred realistic body posture and hand gestures while talking in the communicative scenario to improve their performance. However, there was no significant difference (p>0.05) between these two conditions for verbal behavior ($t(24) = 0.9, p = 0.37$), because the method used to control mouth movements was the same in both approaches.

5.3.3 User Preference

The subjective opinions about ease of use and system preference can be found in Figure 5.7. The results show that 76% of participants thought it was easier to use a depth-sensor-based avatar control approach. Furthermore, about 84% of participants preferred to use depth-sensor-based avatar control.

Participants also gave some comments about the overall experience.

- *"Depth-sensor-based avatar control is more realistic for mapping my hands in the virtual environment"*
- *"I believe that the depth-sensor-based avatar control system is comparatively much easier, which can give you much more freedom than the controller-based*

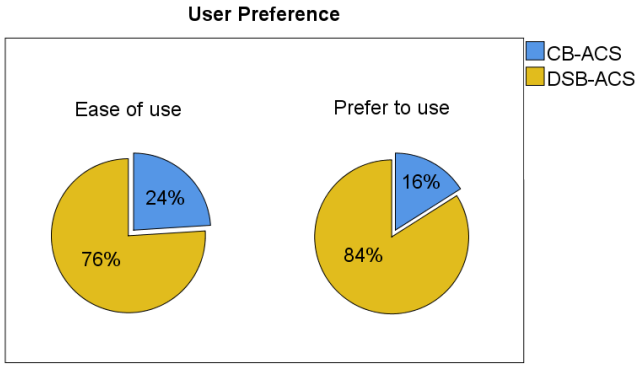


FIGURE 5.7: Usability, Workload and performance

system as you might have the feeling that you are possessing something in your hands. The use of an object makes you feel that you are not truly connected to the virtual world"

- *"I feel completely immersed in the depth-sensor-based avatar control system"*
- *"The depth-sensor-based system shows hand movements and finger movements more realistically and I could move my hands easier in the interview"*
- *"The posture looked more real in the depth-sensor-based system, and tracking of arms was better too"*

5.4 Discussion

In this experiment, hypotheses H_1 and H_2 were rejected as there were no significant differences between the two avatar control systems on the sense of presence and mental workload. The results were unexpected. I thought participants might feel a greater presence using Depth-sensor based avatar control system, as it provided a more natural way for interaction. I did, however, find support for hypotheses H_3 and H_4. The subjective questionnaire responses showed a significant effect that the depth-sensor-based avatar control approach elicited a higher sense of virtual body ownership illusion and agency, as well as better usability compared to the controller-based avatar control system. In support of hypothesis H_5, It was found that participants indicated that they had better performance in the communication scenario

using the depth-sensor-based avatar control system in terms of non-verbal behavior cues, but not in terms of verbal performance between these two systems. Moreover, when asked for a preference, most participants indicated that they preferred the depth-sensor-based avatar control approach and that it was easier to use.

Based on these results, I would suggest that VR developers should adopt methods for full-body tracking that are as expressive as possible. Depth-sensor-based hand tracking can provide an intuitive way to support gesture-based interaction, and users do not always understand button or trigger mappings. The animation recording and reply mode provided a way to make self-evaluations from a third-person perspective; it is not only more flexible compared to inviting another person into VR, but also provides an objective way to review performance compared to facing a virtual mirror, especially for training and single-user communication simulation systems.

5.4.1 Limitations

In this user study, I found some technical limitations. First, the hand-tracking data sometimes switched between the fused Kinect system and the Leap Motion system, and there was no finger data when the participants moved their hands outside the tracked area of the Leap Motion. Some participants noticed a slight hand pose change between the tracking boundaries in the interview review part. I believe that the subtle pose change did not have an impact on the experiment since only three participants reported it. Second, the first two participants reported body penetration effects while they reviewed their virtual interviews. The arms slightly penetrated the body, which was because the participants were nervous during the interview, and put their arms too close to their bodies. This occlusion issue can cause bad recognition from Kinect sensors. Therefore, participants were asked to be relaxed and to keep some space between their arms and bodies. These two cases only happened during the interview question part, and there was no such issue during the route-planning task. I do not think there was any impact on the results since

there were no similar problems reported by the rest of the participants. In this user study, there may be confounding between tracking performance and hands-free interaction. The system can provide realistic and natural hand interaction. Maybe this advantage outperformed the controller-based avatar control system because of the tracking performance. In a future user study, we should consider changing the tracking performance and compare the different levels of hands-free interaction.

5.5 Conclusions and Future Work

In this chapter, the effects of a depth-sensor-based avatar control approach on presence, virtual body ownership, mental workload, usability, and communication behavior were investigated. An avatar control method was provided that supports realistic behaviors based on data from multiple depth sensors (multiple Kinects and a Leap Motion). The fully-tracked body and hand-gesture avatar control system was compared to a controller-based IK system as a baseline condition. I found significantly higher virtual body ownership illusion and usability as well as better non-verbal communication performance by participants in the depth-sensor-based experience compared to the controller-based experience.

However, the limitation of a single Leap Motion was found in this system. In order to address the problem, the development and impact of a multi-Leap-Motion-Controller system will be introduced for providing an extensive and usable hand tracking experience in the next chapter.

Chapter 6

Robust Hands Tracking with Enlarged Tracking Area

The Leap Motion Controller (LMC) is a widely-used 3D user-interface device for hand tracking applications, and is also widely used in VR applications. A LMC was integrated in the body tracking system described in Chapters 4 and 5. However, the tracking area of a single LMC is not sufficient to cover the complete range of hand motions, which can cause inconvenience and unnatural behavior of bare-hand interaction in a collaborative virtual environment. In this chapter, fusing the data from multiple LMCs is proposed to enlarge the tracking area. The configuration of the five-LMC system used on an Oculus Rift S was firstly described. Then, the shared-view calibration method based on the Least-squares fitting algorithm was discussed. To avoid incorrect tracking data from a single LMC interfering with the fusion result, a multi-LMC fusion algorithm based on two-level data evaluation was proposed, which consists of a prediction-based and a position-based evaluation method. Based on the evaluation result, the data was combined from multiple LMCs using a Kalman Filter sensor fusion approach. The system experiment shows that my system can enlarge the hands tracking range to 202.16 degrees horizontally and 164.43 degrees vertically. Then the system performance and the tracking stability was discussed, even in the presence of outliers. The contribution of this chapter is to provide a detailed guide for designing an enlarged hand-tracking system using sensor fusion. This work was submitted as a full paper in the IEEE Sensors Journal.

6.1 Introduction

In Chapter 5, it was found that using a marker-less body tracking system with hand gestures showed better usability than hand-held controllers, which negatively interfered with natural body movements. Proposed solutions from the research literature include the use of multiple depth cameras [138] for more-accurate gross motor tracking, the use of short-range cameras such as the LMC for fine motor tracking of the hands and fingers [96], and a combination of both [140]. However, the narrow field of the tracking area was a problem, since the LMC tracks hand movement only in the front area of the user's head according to the single LMC settings in VR mode. To perform interaction in other areas, such as the lateral regions of the body, the user needs to rotate their head to the interested area and look at the hands all the time, which causes unnatural behavior and inconvenience in collaborative VEs. For example, if two people shake hands in a virtual environment, they need to stare at their hands to avoid having the LMC lose track of them, rather than looking at the face of the other. To solve these issues, multiple LMCs solution was proposed to extend the hand-tracking area.

The challenge of extending hand tracking by combining data from multiple LMCs is how to eliminate erroneous tracking data. The tracking accuracy of the LMC is easily affected by the ambiguity of depth data and perspective distortion, which cause erroneous tracking results of the wrong hand and/or inaccurate tracking (Figure 6.1). The LMC SDK has two modes of operation, Desktop-optimised and HMD-optimised, and the errors often occur when the tracking environment does not match one of these. However, the tracking environment of any additional LMCs in the multi-LMC system poorly fits either of these modes, which will generate incorrect tracking data. Therefore, the data confidence of additional LMCs should be evaluated before any fusion process. The confidence parameter provided by the LMC SDK only estimates the confidence of the gesture recognition, and cannot be used to represent the reliability of tracking data. To address the problem, Jin et al. [59] fused the tracking data from two LMCs by evaluating the tracking status of every single

finger, but their fusion algorithm could not deal with erroneous tracking data. Hu et al. [55] used a Markov method to fuse the data from five LMCs, but they did not present the details of their fusion algorithm. Besides, the tracking area of LMCs in Jin and Hu's system was fully overlapped with each other, which did not enlarge the tracking area of the system. Feuchtner et al. [34] provided a partial solution for tracking hand movement in the front and lower front of the body from an HMD. However, the details of the calibration were not presented, and they did not combine data from multiple LMCs. Instead, the system switched between the two LMCs when the hand passed from one device's tracking area to another, keeping only one LMC active. Hence, it would not be possible, for example, to track two hands in different tracking areas at the same time. Thus, a data fusion algorithm that can simultaneously track two hands in the enlarged tracking area is required.

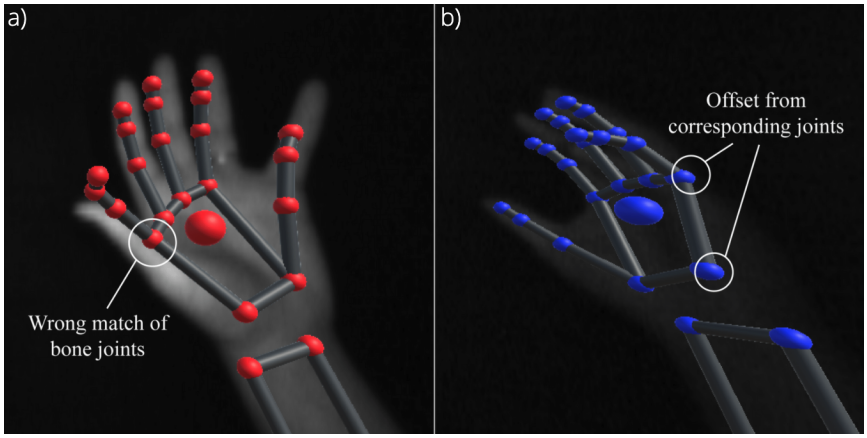


FIGURE 6.1: Examples of erroneous tracking results. a) Wrong-hand detection due to the ambiguity of depth data. The thumb of the virtual hand is wrongly aligned with the pinky of the real hand. b) Inaccurate tracking data due to high distortion at the edge of the LMC's tracking field leads to an offset between the joints of the virtual hand and the real hand.

In this chapter, a comprehensive approach is presented by using multiple LMCs to enlarge the hand-tracking area for VR applications. The setup of a five-LMC system used with the Oculus Rift S, and an efficient calibration method based on the Least-squares fitting (LSF) algorithm were described. Then a multi-LMC data fusion algorithm was introduced that uses a two-level method to evaluate the tracking performance of a single LMC and

combines the data from multiple LMCs based on the evaluation results. The contributions of this work are:

- An evaluation method of tracking data confidence based on the skeleton data output by a single LMC.
- A multi-LMC data fusion algorithm that can enlarge the hand tracking area by 34% in the horizontal and 37% in the vertical area while accurately tracking two hands in the enlarged tracking area.
- A prototype that provides a detailed reference for designing an enlarged hand-tracking system using multi-sensor data fusion in VR.

The Oculus Quest[1] has four ultra-wide cameras for large area hands tracking. But the software is the beta version, and the link mode is not available since it was released. It would be better to compare our multi-LMC system with the Quest for robust and detection range in the near future.

6.2 System

In this section, the hardware and software in this system is described. The configuration and design of the multi-LMC mount with the design approaches will be presented.

6.2.1 Hardware

Installing the additional LMCs on the static object around the users [59] limits the interaction area and increases the calibration difficulty. Attaching the LMCs on other parts of the body than the head also brings problems of real-time calibration of the relative position between two LMCs on two different body parts. Thus, I propose to integrate all LMCs on the HMD helmet. Figure 6.2 shows the physical configuration. Five LMCs are used in the system. The central LMC is attached in the middle of HMD for capturing hand movement data in front of the user. The lateral LMCs at the four corners

[1] https://www.oculus.com/

of the HMD provide supplementary tracking in the top-left, top-right, bottom-left, and bottom-right areas. The lateral LMCs are positioned relative to the observing coordinate system, whose origin is located at the center of the front surface of the HMD with the x-axis facing left, the y-axis facing up, and the z-axis facing forward. According to the maximum position that the human hand can reach [78], the positioning parameters of the four lateral LMCs are presented in Table 6.1. The parameters ensure that the tracking area is large enough to cover the whole hand movement range, while keeping the overlapping areas to be sufficient for calibration. The error caused by infrared interference in the configuration is negligible [95].

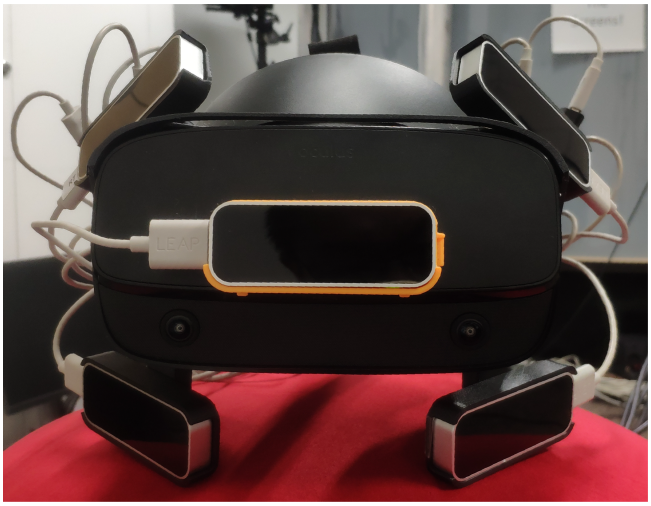


FIGURE 6.2: Multi-LMC mount on the Oculus Rift S

Due to USB bandwidth limitations, it is not possible to simultaneously connect five LMCs to a single computer. In the system, the front LMC is connected to the main workstation (Intel Core i7-7700K at 4.2 GHz, 32GB RAM, and NVIDIA GeForce 1080Ti), which also drives the Oculus Rift S and the VR scene in Unity. Four Intel NUC computers (Intel Core i5-8259U at 2.3 GHz, 8GB RAM, and Iris Plus Graphics 655) are deployed for connecting the four lateral LMCs. The data between the NUCs and the workstation is transmitted based on the UDP protocol within a local area network through a gigabit switch (NETGEAR GS110MX). The data-transmission latency between the NUCs and the main workstation is less than 2ms. The bandwidth requirement

is 2.4Mb/s for single-hand data and 14.4Mb/s at peak. The rendering rate of the Unity scene is higher than 30Hz with the setup.

TABLE 6.1: Positions and Rotations of four side leap motion

LMC position	**Translation (mm)**			**Rotation (degree)**		
	x	y	z	x	y	z
Top-left	80	50	-60	-35	35	-30
Top-right	-80	50	-60	-35	-35	30
Bottom-left	80	-75	-80	30	35	30
Bottom-right	-80	-75	-80	30	-35	-30

6.2.2 Software

The primary system is built using Unity version 2019.2.0f1 with the Leap Motion plugin (Core 4.4.0) [74]. The Point Cloud Library (PCL) [85] version 1.6.0 is used to perform ICP calculations during the calibration process. A hand-tracking data collection and serialization application were built on the NUCs based on the Leap Motion SDK (4.0.0) [94]. The data transmission between the workstation and the NUCs is achieved using the Rug.OSC library [104].

6.3 Method

This section introduces the calibration method and the fusion method used in my system. The data flow and process are shown in Figure 6.3.

6.3.1 Calibration

The built-in re-calibration function was used to calibrate the intrinsic parameters of the individual LMCs in my system. As for extrinsic calibration, an efficient approach is proposed to calibrate the multi-LMC array with no dependence on external devices. Because the overlapping tracking range of the LMCs is sufficient for calibration, a shared-view method based on the LSF algorithm was devised to calibrate multiple LMCs.

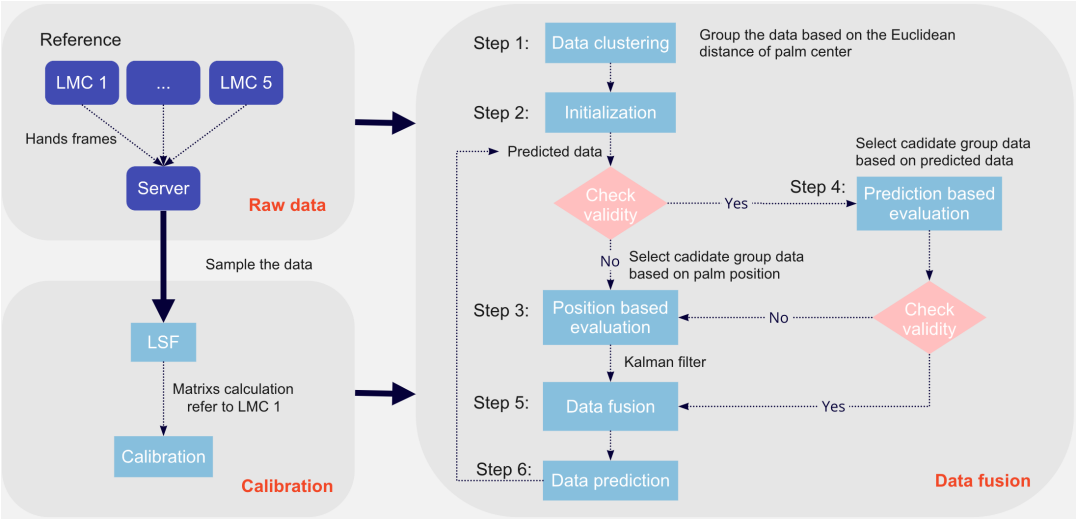


FIGURE 6.3: Multi-LMC data flow chart

The front LMC was set as the reference camera. The input is the hand trajectories of the specified hand-joint sampled by the reference LMC and the lateral LMCs in the overlapping tracking area. During sampling, the user needs to flatten their hands and move the hands randomly in the overlapping tracking area (left hand in the overlapping area of central, top-left and bottom-left LMCs, right hand in the overlapping area of central, top-right and bottom-right LMCs). In order to eliminate the error caused by the sampling latency between each LMC, the moving speed of hands should be slow (less than 10 millimeters per second according to my experience). C_r was used to represent the trajectory from the reference LMC, and use $C_s, s = 1, 2, 3, 4$ to represent the trajectories from the lateral LMCs. After sampling, two point sets $P_r = \{p_r | p_r^{(i)} \in C_r, i = 1 \dots n\}$ and $P_s = \{p_s | p_s^{(j)} \in C_s, s = 1 \dots 4, j = 1 \dots n\}$ are generated from C_r and C_s, respectively.

The next step is to calculate the calibration matrix using the LSF algorithm. The theory of LSF is to find the optimal transformation (consisting of rotation R and translation t), which minimizes the sum of the distance between the coordinates of the matching pairs [143]. R_s and t_s are used to represent the transformation parameters of the s-th lateral camera. The objective function of the LSF algorithm can be represented using Equation 6.1.

$$f(R_s, t_s) = \sum_{i=1}^{n} \|R_s x_i + t_s - y_i\|^2, x_i \in P_s, y_i \in P_r \tag{6.1}$$

in which x_i and y_i are a pair of corresponding points between P_s and P_r. Because my method samples the data from the reference camera and the calibrating camera simultaneously, the corresponding-point pair in my method is given by Equation 6.2.

$$(x_i, y_i) = (p_s^{(i)}, p_r^{(i)}), i = 1 \dots n \tag{6.2}$$

Substituting Equation 6.2 into Equation 6.1, the objective function then becomes Equation 6.3.

$$f(R_s, t_s) = \sum_{i=1}^{n} \|R_s p_s^{(i)} + t_s - p_r^{(i)}\|^2 \tag{6.3}$$

Equation 6.3 can be solved using the singular value decomposition (SVD) method [4]. In my system, the software automatically ran the SVD solver from the PCL library to calculate the calibration matrix after the sampling. According to my pilot test, an experienced user can complete the entire calibration process in 2 minutes. Once calibration is complete, there is no need to recalibrate unless the LMCs are moved.

6.3.2 Multi-LMC Data Fusion

The main task of my fusion algorithm is to find the most reliable data set, named candidate group g_c, from the raw skeleton data provided by the LMC SDK. Then, the algorithm combines the data based on the data confidence μ. A vector $\mu_c = \{\mu_c^l, \mu_c^r\}^T$ is introduced to ensure the chirality correctness (handedness) of the fused hands. The value of μ_c^l and μ_c^r indicate the confidence of the left or right chirality expressed by the group of data.

Algorithm Overview

An overview of the steps of my algorithm is shown below.

- **STEP 1: Data Clustering** - This step firstly clusters the data from the raw data set, which contains all the detected hand data from multiple LMCs, into a group set $G = \{g_1, \ldots, g_j\}$ according to the palm center position. The group $g_j = \{h_1, \ldots, h_k\}$ is a collection of tracking data of hand h from k LMCs. The Euclidean distance between the palm center of h_k in the group g_j is within a threshold ϵ. The calibration error determines the value of ϵ.

- **STEP 2: Initialization** - The validity of the predicted data h_p made in the last frame is checked. If h_p is valid, the algorithm will go to **STEP 4**. Otherwise, the algorithm will go to **STEP 3**.

- **STEP 3: Position-based Evaluation** - The data confidence μ of all detected hand data is calculated based on the palm center position. Then, the μ_c of each group in G is calculated using the evaluation result. After that, the candidate groups are selected out by comparing μ_c among all groups and sent to **STEP 5** for data fusion.

- **STEP 4: Prediction-based Evaluation** - First, the groups closest to h_p are chosen as the candidate groups. Then, the μ of hands in the candidate group is calculated based on the skeleton data difference between the tracking data and the predicted data. Finally, the evaluation results are verified using a chirality verification method. If the result is valid, the algorithm will go to **STEP 5**. Otherwise, the algorithm will go to **STEP 3**.

- **STEP 5: Data Fusion** - The fused results are obtained by fusing the hand data in the candidate groups according to the confidence μ_c. The chirality of the fused results is decided according to the u_c. The data of the candidate group will be fused with h_p using a Kalman filter if h_p is valid.

- **STEP 6: Prediction** - If the last frame data is valid, the hand motion of the next frame will be predicted based on kinematic theory (described

in detail below). Then, the fusion data of the current frame is stored for the prediction process of the next frame.

More details of my algorithm are given in the following parts.

Position-based Evaluation

The theory of the position-based method is based on the inconsistency of LMC tracking quality [49], which considers that the hand-tracking quality will be good if the hand is close to the center of its observing LMC's tracking range. The function of the confidence calculation is given in Equation 6.4:

$$\mu^{position} = \frac{\epsilon_a}{\epsilon_a + d_c} \tag{6.4}$$

in which d_c (mm) is the distance between the palm center of the detected hand and the y-axis of the observing LMC's coordinate system, and ϵ_a is an empirical parameter which represents the range of good tracking quality. In this chapter, ϵ_a was set to 250mm according to Joze's work [49]. The μ_c of each group is calculated using Equation 6.5:

$$\mu_c = \sum_{i=1}^{k} \mu^{(i)} \mu_c^{(i)} \tag{6.5}$$

where $\mu_c^{(i)}$ is the chirality confidence of each hand in the group, acquired from the estimation result of the LMC SDK. $\mu_c^{(i)}$ equals $(1, 0)^T$ if the hand is estimated as a left hand or equals $(0, 1)^T$ if the hand is estimated as a right hand. The group with the highest value of μ_c^l is selected as the candidate group for left-hand fusion. For μ_c^r, the rule is the same for the right-hand fusion.

It should be noted that the position-based evaluation method is a rule-of-thumb method. The result of the confidence evaluation will be unreliable sometimes. However, this method does not require the data of previous frames. Thus, it is used to calculate the initial value for the prediction-based method and as a supplementary method when the prediction-based method generates invalid results.

Prediction-based Evaluation

The theory of the prediction-based method is based on the spatio-temporal continuity of hand motion [2], which considers the data difference between the current frame and the prediction from the last frame to be smaller for the correct tracking data compared to poor tracking data. Because the four metacarpal bones of the palm can be regarded as a rigid body, it is reasonable to predict the motion of these bones using the palm center velocity. Thus, the position of the *PrevJoint* and *NextJoint* of the four palm metacarpal bones are chosen to calculate the data difference. The function of the prediction-based evaluation is given as Equation 6.6:

$$\mu^{prediction} = \frac{\epsilon_p^3}{\epsilon_p^3 + d_m^3} \tag{6.6}$$

in which ϵ_p is the indicator of 50% confidence and d_m is the sum of the distance of the metacarpal joint between the tracking data and the predicted data, which is calculated as Equation 6.7:

$$d_m = \sum_{i=1}^{8} \sqrt{(p_t^{(i)} - p_p^{(i)})^2} \tag{6.7}$$

in which $p_t^{(i)}$ and $p_p^{(i)}$ are the position vector of the metacarpal joints of the tracking data and the predicted data, respectively.

The data confidence indicator ϵ_p is related to the distribution of d_m under the normal and poor tracking conditions. To ensure a safe classification, I choose the mid-value between the upper-bound and the lower-bound of the 99.7% confidence interval of the normal and poor tracking distribution respectively as ϵ_p.

A verification process was used to ensure the correctness of the evaluation result because the prediction results are not reliable when the hand moves quickly. In the process, the algorithm compares the μ_c^l with μ_c^r for each candidate group and chooses the larger one as the chirality according to the evaluation result. If the chirality of the evaluation result is coincident with the prediction, the evaluation of this group is considered as reliable. Otherwise,

the result will be discarded and use the position-based method to evaluate the current frame.

Data Fusion

After acquiring the data confidence of each hand from the above evaluation method, a weighted-sum method is used to obtain the fused result of the candidate group g_c. The function is presented as Equations 6.8 and 6.9:

$$h_f = \sum_{i=1}^{k} \omega_i h_i \tag{6.8}$$

$$\omega_i = \frac{\mu_i}{\sum_{j=1}^{k} \mu_j} \tag{6.9}$$

Where h_f and h_i represent the skeleton joint pose of the fused hand and original hand respectively, and ω_i is the weighting value of h_i. Because the difference of the rotation data between the hand in g_c is small after calibration, a linear method was used to calculate quaternion interpolation approximately.

A Kalman filter was used to improve the fusion quality if the prediction data is valid. Assuming that the tracking error of all hand joints follows the same distribution, the update function of the Kalman filter [10] can be given as:

$$h'_f = h_p + K(h_f - h_p) \tag{6.10}$$

$$P' = (1 - K)P \tag{6.11}$$

$$K = \frac{P}{P + R} \tag{6.12}$$

In the above equations, h'_f is the final fusion result of the current frame, and P' and P are the variance of the final fusion results and the prediction results, respectively. K represents the Kalman gain. R represent the variance of the fused tracking data of the current frame. Because the calibration accuracy

affects the absolute position measurement of each LMC, R can be represented by Equation 6.13:

$$R = \sum_{i=1}^{n} \mu_i^2 R_i \tag{6.13}$$

in which R_i is the calibration error of each LMC.

Prediction

The prediction of the hand motion in the next frame is based on kinematic rules. The velocity of the hand v_t was calculated firstly using Equation 6.14:

$$v_t = \frac{p_t - p_{t-1}}{\Delta_{t-1}} \tag{6.14}$$

in which p_t and p_{t-1} are the palm center position of the current and previous frames, respectively, and Δ_{t-1} represents the time interval between the current and previous frames. Then, the predicted result of next frame is obtained by Equation 6.15:

$$h_{p,t} = h'_{f,t-1} + v_t \Delta_t \tag{6.15}$$

in which, Δ_t is the time interval between the current frame and the next frame.

6.4 Experiment

Two tests are presented in this section. In the first test, the system parameters were tested, including calibration error, prediction error, tracking range, and distribution of d_m under normal and erroneous tracking conditions. Then, the performance of the fusion algorithm was examined in the enlarged tracking area.

6.4.1 System Parameter Test

The system was setup and calibrated each lateral LMC using the LSF-based calibration method described above. The number of sampling points for

calibration was set to 300, and the data of the palm center joint was sampled to calculate the calibration matrix.

For measuring calibration error, the tester was asked to place the hands statically in the center of the overlapping area with fingers opened. Then, the data was sampled from the five LMCs.

For measuring prediction error, the center LMC was chosen as the testing device. During the test, the tester was asked to randomly move his/her hand in the tracking area of the center LMC, and recorded the tracking data of the current frame and the prediction data calculated by Equation 6.15.

In the test of d_m distribution, the center, bottom-left, and bottom-right LMCs were used to sample data. Erroneous (poor) tracking conditions were simulated by setting the tracking policy of the bottom-left LMC to desktop-optimized mode. The policies of the center and bottom-right were set to HMD-optimized mode to provide reference data and normal tracking data, respectively. During the test, the tester was asked to move the left hand in the left-side overlapping area and right hand in the right-side overlapping area simultaneously. The data of the three LMCs were recorded at the same time.

In terms of data sampling, the measurement parameters of each test were sampled 1000 times, and the position of the *PrevJoint* and the *NextJoint* of the four palm metacarpal bones were recorded to calculate the d_m. One tester was invited to perform my experiment in slow, medium, and fast speed, and the average was chosen as the test result.

After sampling, Equation 6.7 was used to calculate the error or difference of each test. The calibrating error of each lateral LMC was obtained by calculating the d_m relative to the center LMC's data. The prediction error is the d_m between the current tracking data and the prediction results. By using the bottom-left and bottom-right LMC's data to calculate d_m with the center LMC's data, the d_m distribution was accquired under erroneous and normal tracking conditions. After the calculation, a statistical analysis of each test result was performed.

Finally, the tracking range was tested by waving the hands 10 times horizontally and vertically and recording the hand palm position while waving.

The tracking range was obtained by calculating the maximum range that the hand data could reach.

6.4.2 Performance Test

The primary goal of the Multi-LMC system is enlarging the usable hand tracking area while maintaining a smooth data transition when hands pass through overlapping tracking areas and reducing the effect of erroneous tracking data. Therefore, the performance test aims to compare the tracking range of my system with a single LMC and test the system's ability to dispose of outliers. Besides, the small offsets between the real hands and the virtual hands in virtual environments have little effect on users' feeling of presence [83]. Thus, I did not compare the system measurement with the ground truth. To test the performance in the enlarged tracking area, a moving-box task was designed in a virtual environment using Unity. The task structure is shown in Figure 6.4. The user firstly fetched a box at the lower-left area of the body using his/her left hand. Then, the user needed to move the box to pass through seven points around his/her body in sequence. The path points were distributed relative to the center LMC of the multi-LMC system. A virtual mirror was placed in front of the user to provide a clear vision of the box moving. An interaction panel was designed in the virtual scene to help the user perform the calibration and data collection operations.

The system was set up according to the configuration in Figure 6.4, and set the optimization policy of the five LMCs as HMD-Optimized. Before the test, the user was asked to calibrate the system three times. The result with the minimum calibration error was chosen to calibrate the system. When the calibration was ready, the user activated the data collection function using the UI panel, and a 10-second countdown was initiated to get the user in place. Then, the user started to perform the task and was asked to keep looking forward while the task is running. The trajectory of the user's left and right hands, as well as the confidence values of the five LMCs, were collected.

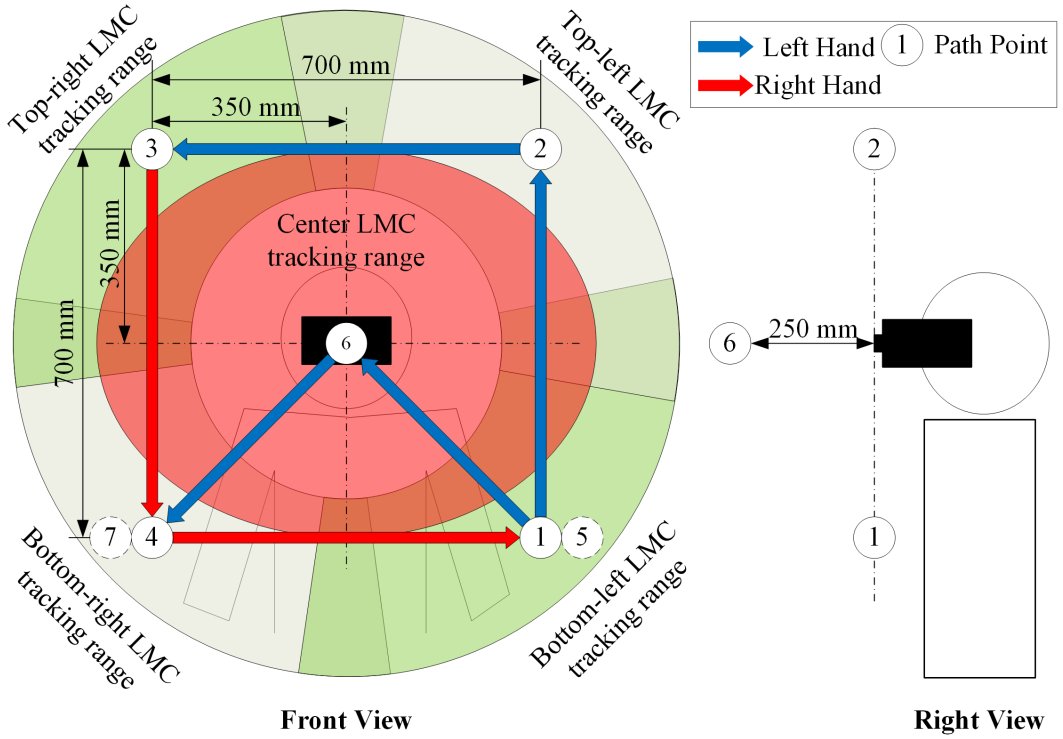


FIGURE 6.4: The task setting in the performance experiment

When the task was completed, the user accessed the UI panel to stop the data collection function.

According to the sketch of each LMC's tracking range shown in Figure 6.4, the detected hand should pass through different types of tracking regions, including standalone, double-overlapping, and triple-overlapping tracking. The reason I chose the transporting box as the task is that the grabbing gesture has the lowest tracking performance of the LMC. Thus, the task could represent the worst tracking condition.

6.5 Results and Discussion

6.5.1 System Parameters

Table 6.2 shows the statistical results of calibration error, prediction error, and the joint difference d_m under normal and poor tracking conditions. Figure 6.5 illustrates the comparison of the mean and the standard variance of these data. From the results, I found a clear difference between the mean value

of d_m under the two tracking conditions, which demonstrates the rationality of using the joint position difference as the measurement of data confidence. I found a discrepancy between the calibration error and d_m in the normal tracking condition. The discrepancy between the two data is caused by the latency of data transmission between LMCs. The latency causes a slight offset between the hand data from the lateral LMCs and the data from center LMC. However, the offset caused by the system latency is trivial compared to the error caused by poor tracking results. The prediction error comes from the irregular hand movement; that is, the prediction error is large when the hands are suddenly turned. Thus, the prediction result could be unreliable when the hands were moving quickly.

Error/Difference	Mean (mm)	SD (mm)	kurtosis	skewness
Calibration	34.10	3.24	-0.95	0.15
Prediction	28.91	17.55	10.18	2.14
Normal tracking	83.17	29.86	3.13	1.41
Poor tracking	407.47	35.62	1.35	0.88

TABLE 6.2: Descriptive statistics of the error and difference distribution in the system parameter experiment

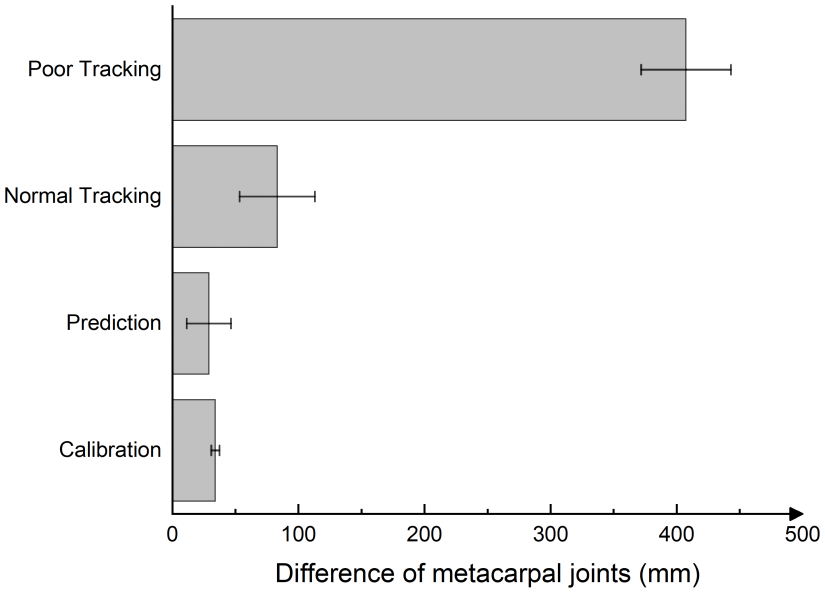


FIGURE 6.5: Bar-chart of the descriptive statistics

Based on the above results, the good system parameters can be determined. The mean prediction error is used as the initial value of P in Equations 6.11 and 6.12, and the mean calibration error is assigned to R in Equations 6.12 and 6.13. The value of ϵ_p is set to 211.56mm according to the definition in Equation 6.6.

Table 6.3 shows the comparison of the available tracking range between my system and a single LMC in horizontal and vertical directions. The results show that, compared to the official single LMC tracking range information, my system enlarges the horizontal tracking range to 202.16 degrees, an increase of 34%, and the vertical tracking range to 164.43 degrees, an increase of 37%.

Tracking Range		Multi-LMC (degrees)	Single LMC (degrees)
Horizontal	left edge	-101.07	-75
	right edge	101.09	75
	Range	202.16	150
Vertical	upper edge	89.93	60
	lower edge	-74.50	-60
	Range	164.43	120

TABLE 6.3: Comparison of tracking range between the multi-LMC system and single LMC

6.5.2 Fusion Performance

Figure 6.6 shows the trajectories of the left and right fused hands in the moving-box task. The purple and dark blue traces show that the left hand first passed through Points 1 and 2 and reached Point 3, while the right hand was waiting at Point 4. Then, the right hand moved to Point 3 to fetch the box, and the left hand moved to Point 5, which is shown in the blue and light blue traces. After that, as the light blue and green traces show, the right hand passed through Points 4 and 5 and moved back to Point 4 after the box was transferred to the left hand, while the left hand was waiting at the Point 5. Finally, the yellow and orange traces show that the left hand passed through Points 6 and 7 to finish the rest of the task. The red trace represents the path

that left and right hand moved to the center to stop the data recording, the bottom-left LMC wrongly recognized the left hand's chirality at this time. The trajectory continuity and the stable fusion results of hand chirality reveal that my algorithm can accurately fuse the data from multiple LMCs.

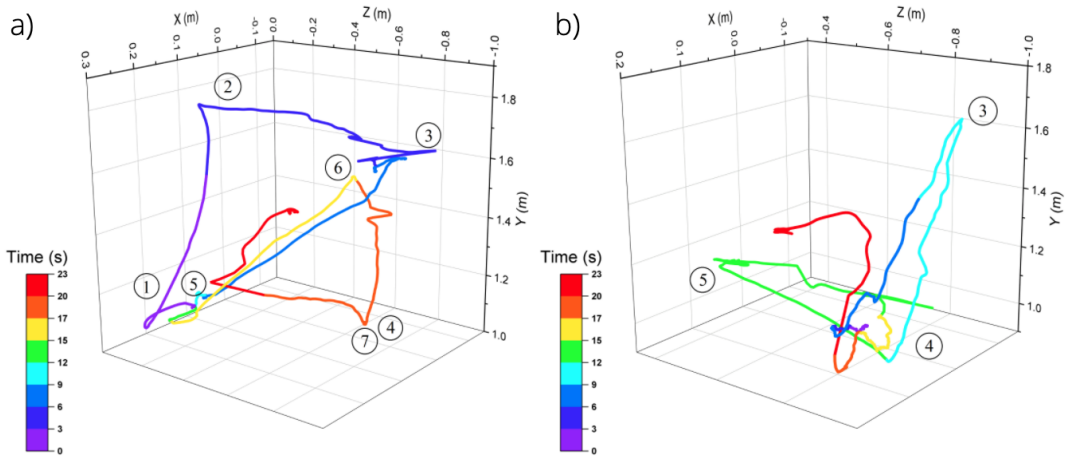


FIGURE 6.6: Trajectory of fused hands in the performance experiment. The trajectories of the left and right hands are continuous in different tracking regions. a) Left hand trajectory, b) Right hand trajectory

Figure 6.7 presents the confidence values of each LMC evaluated by the prediction-based method and the position-based method. The figure clearly shows the oscillation of confidence values when the hands cross the tracking border between two neighboring LMCs. This might be the result of the LMC's unstable tracking for the suddenly emerging hands. Compared with the position-based method, the confidence values evaluated by the prediction-based method are more decentralized, which reveals that the prediction-based method performs better at suppressing unreliable data. In contrast, the result of the position-based method changes less dramatically than that of the prediction-based method.

The confidence results coincide with the expectation of the experiment task, except for a wrong tracking case marked in Figure 6.7(d). During the period of the wrong tracking case (20.5s to 22s), the user moved their left and right hands from the sides to the front in order to use the UI panel to stop recording data. In the ideal condition, the LMCs on the left side would only collect data from the left hand, while the LMCs on the right side would only

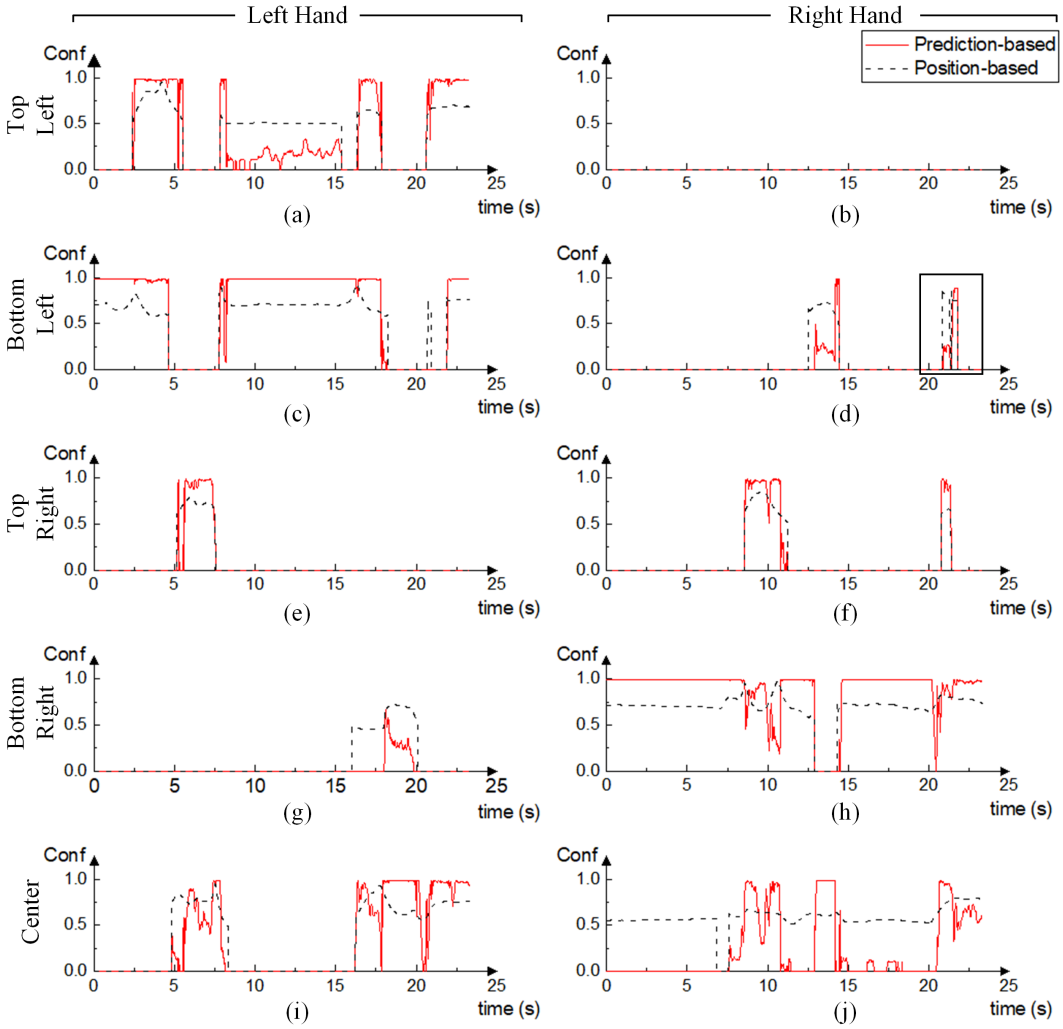


FIGURE 6.7: Confidence evaluation results of the prediction-based method and the position-based method in the accuracy experiment. Each row of sub-figures represents the hand confidence of the LMC at different positions. Each column represents the confidence of the left and right hands in each LMC. The solid red line represents the confidence evaluated with the prediction-based method, and the black dashed line represents the confidence evaluated with the position-based method. The rectangle in (d) marks a set of incorrect tracking data from the bottom-left LMC, which wrongly recognized the left hand as the right hand.

sample the data of the right hand. Only the center LMC has the opportunity to collect the data from both hands. However, the marked area in Figure 6.7(d) shows that the bottom-left collected the data from the right hand during this period, which means the bottom-left LMC wrongly recognized the left hand as the right hand and generated incorrect confidence values. Figure 6.8 shows the evaluation results of the prediction-based and position-based methods in this period. From the figure, I found that the prediction-based method performs better than the position-based method in terms of the ability to recognize and suppress wrong tracking data. The average weighting of the wrong tracking data (right-hand data from bottom-left LMC) in this period is 0.19 for the prediction-based method and 0.33 for the position-based method. For the prediction-based method, the confidence values of the wrong tracking data are lower than that of the correct tracking results, and the weighting calculated using the prediction-based method is lower than the weighting calculated using a position-based method. However, there was no significant difference in the evaluation results made by the position-based method. This is because the prediction-based method can sensitively identify the erroneous tracking data according to statistical characteristics. However, the position-based method empirically evaluates the confidence according to the hand position, which assigns similar confidence values to the poor tracking data when the hand is at an intermediate point between multiple LMCs.

As for robustness, the prediction-based method is revealed to have less stability than the position-based method by showing more fluctuations in the evaluation results and several instances of failed evaluation, such as the missing value in the period from 16s to 20s in Figure 6.7(g). This is because the prediction-based method relies on initial values, and may sometimes fail to evaluate the confidence when the prediction process is unstable. Thus, the prediction-based method needs the position-based method to provide the initial value and works as an alternative method when errors occur.

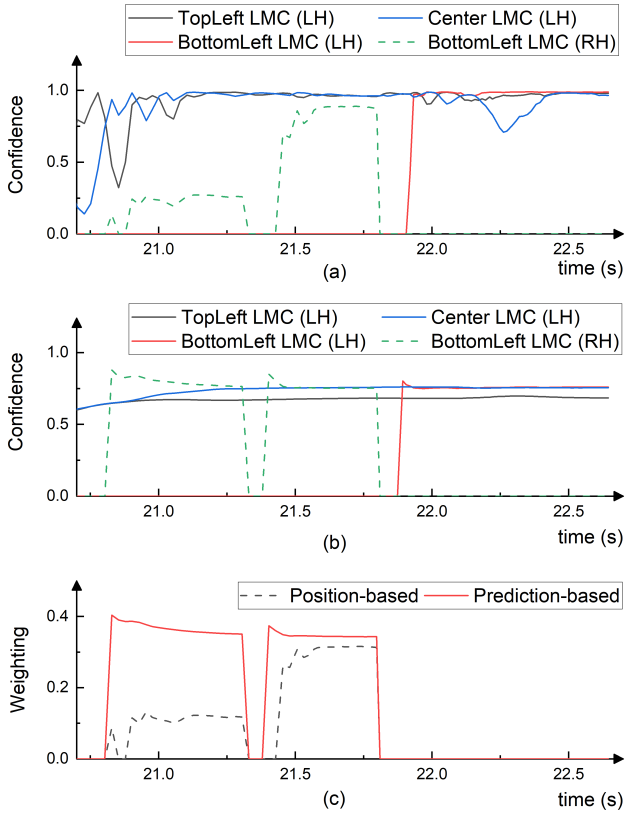


FIGURE 6.8: The confidence and weighting in the wrong tracking case. (a) is the hand confidence of top-left, bottom-left and center LMC evaluated by the prediction-based method. (b) is the hand confidence of the three LMCs evaluated by the position-based method. (c) is the weighting of the wrong tracking data calculated by Equation 6.9 using the evaluation results of the position-based and prediction-based methods. LH and RH are the acronyms of left hand and right hand, respectively.

6.5.3 Limitations and Scalability

Current limitations of system accuracy mainly come from the measuring accuracy of a single LMC. I found that a hand-shifting phenomenon happens when the user holds their hands forward and turns the head around. This leads to uneven calibration error in the tracking field of my system, which weakens my ability to estimate the measurement deviation accurately. The reason comes from inconsistencies in the LMC tracking performance. Therefore, it is necessary to calibrate the intrinsic parameters of each LMC.

The prediction error is another factor that affects the performance of my system. Due to system latency and packet loss, the prediction results could sometimes be unreliable, resulting in errors in the fusion results. Better network connectivity (low latency) and some restrictions (constrains on prediction results) would reduce the impact of unreliable prediction results.

In terms of scalability (i.e., adding more LMCs), the main limitation is the high demand for data transmission bandwidth. In my observations, the bandwidth needed for transmitting a single hand is 2.4Mb/s. A system with five LMCs requires a bandwidth of 14.4Mb/s at peak, which is a heavy network load. Also, since each LMC needs its supporting computer, the number of LMCs is limited by the number of additional devices and cable mobility.

6.6 Conclusions and Future Work

This chapter explored the feasibility of using multiple Leap Motion Controllers (LMCs) to extend the usable tracking area for hand-based interactions in VR. A five-LMC system was built and attached to an Oculus Rift S HMD.

An LSF-based calibration method was introduced. To eliminate the impact of poor tracking data on fusion results, a multi-LMC fusion algorithm based on a two-level evaluation method was proposed. This method is composed of a prediction-based method and a position-based method. The prediction-based method evaluates the data confidence of a single LMC based on the difference

of joint positions between the current tracking data and the prediction from the previous frame. The position-based method uses the relative position between the hand and its observing LMC to evaluate the tracking quality. The data from multiple LMCs are combined according to each LMC's data confidence, and the combined result is fused with prediction using a Kalman filter.

Two tests were designed to assess system parameters and performance. The results of the first experiment illustrate the rationality of using joint differences to measure data confidence. With the setup of my system, the valid tracking range of hand motion can be enlarged to 202.16 degrees horizontally and 164.13 degrees vertically. The result of the second experiment shows that my system can accurately fuse the hand data in the enlarged tracking area. The prediction-based method can significantly eliminate the impact of erroneous tracking data, but it needs initial values and is sometimes unstable. The position-based method has an inferior evaluation accuracy, but is independent of initial values and more stable than the prediction-based method. Thus, the prediction-based method was set as the primary evaluation method and use the position-based method to provide the initial values. When the prediction-based method fails to do the evaluation, the position-based method is used as an alternative method.

My contribution is an evaluation method for tracking data confidence of a single LMC and a multi-LMC data fusion algorithm that can accurately track two hands in an enlarged tracking area. Our work provides a detailed reference for designing enlarged hand-tracking systems using multi-sensor data fusion. It should be noticed that our work is independent of the sensor's built-in SDK. Thus, our algorithm can be ported to any other platform that provides skeleton data. In the next chapter, I will combine the data from heterogeneous devices (i.e., the multi-LMC system and the multi-Kinect system) to provide a full-body tracking system with a high level of expressiveness in terms of non-verbal cues. Furthermore, I will present a study using the highly expressive avatar control system in a collaborative virtual environment.

Chapter 7

The Effects of a Highly Expressive Avatar Control System on Collaboration Behaviors

The accurate and rich representation of non-verbal behaviors, such as body posture and hand gestures, is critical for the sense of agency and virtual body ownership, and user experience in VR. The study in Chapter 5 demonstrated the positive effects of greater avatar articulation control (using sensor tracking) compared to semi-articulated avatar control (using a controller) for single-user VR. In this chapter, I extend that work and assess the impact of asymmetry in the control of articulation of avatars in multi-user VR. To investigate the impact of different levels of avatar expressiveness on a non-verbal collaboration task, a shared virtual environment with two avatar control systems was implemented, and participants were asked to collaborate using asymmetric control schemes. A charades game was designed to measure copresence, social presence, and interpersonal attraction. The results indicate that participants interacting with highly-expressive avatars report deeper social presence and attraction, and exhibit better task performance than those interacting with partners represented using low-expressive avatars. This work was submitted as a full paper to the ACM CHI Conference 2021 on Human Factors in Computing Systems.

7.1 Introduction

Current VR technology enables people to communicate and collaborate in shared virtual environments (SVEs) from different geographic locations. The quality and efficiency of communication and collaboration in VR depend on several factors, such as virtual environment rendering [80, 43], avatar representation [16], latency [37], and state synchronization [93]. Avatars play an essential role in social VR, and avatar realism is one of the main factors that affect the sense of presence, interpersonal interactions, and copresence [119, 62, 63, 61]. Avatar realism is often used to measure avatar quality, which can be divided into appearance and behavioral realism. Most previous work has been done on visual fidelity [71, 126], and avatar appearance influences interaction in all SVEs [88, 108]. The virtual character represents the user and presents all the verbal and non-verbal behavior from the user in the real world. For communication, humans actively use both verbal and non-verbal behavior for the best representation of their intentions. However, people tend to communicate more through non-verbal behavior [77] during social interaction compared to the verbal channel. Therefore, it is essential to study the impact of non-verbal behavior on communication in VR. Previous research studied some aspects of non-verbal behavior, such as eye gaze [42] and facial expressions [127], which have proven to be important factors in SVEs. Expressive avatar systems (integrating non-verbal behavior, such as body movement, hand gestures, facial expressions, and eye gaze) are limited in current immersive systems due to sensory technologies. Although there is still more work to do on improving avatar appearance and realism [16], the impacts of *expressiveness* of avatars in terms of non-verbal behavior has not yet been systematically investigated in VEs with fully embodied avatars.

In this chapter, a collaborative VR system with asymmetric avatar control approaches is presented, which can support different levels of avatar expressiveness in terms of non-verbal behavior. A charades game was implemented in the SVE with different expressive avatar conditions to measure copresence, social presence, and interpersonal attraction. *"Charades is a game of pantomimes:*

you have to "act out" a phrase without speaking, while the other members of your team try to guess what the phrase is. The objective is for your team to guess the phrase as quickly as possible" [87]. The reason I chose this game is to encourage the participants to perform non-verbal behavior to complete an engaging, collaborative task. The avatar control systems was evaluated with a dyadic user study, investigating performance in terms of accuracy and completion time.

The remainder of this chapter is organized as follows: The system set-up for the avatar control systems and system overview can be seen in Section 7.2. In Section 7.3, the details of the experiment is described. Finally, the discussions of the results, conclusions, and future work are presented.

7.2 Technical Setup

The experimental setup was implemented in a large room with two different physical systems. Participants played the game in each system with asymmetric avatar control connected through the local network. Both participants in each dyad could move freely within a 2m circle, and the tracked movement and gestures were reflected on both virtual characters in the SVE. The details of the avatar system, network architecture, and software will be provided in this section.

7.2.1 Asymmetric Avatar Control Systems

In this experiment, two avatar systems with different levels of expressiveness was adopted.

Highly Expressive Avatar (HEA) Control System

The participants who used this avatar control system could control a highly-expressive avatar representation with a contact-less tracking system.

- **Body Tracking** - The full-body movement came from the data integration of four Kinect v2 devices placed in the corners of the tracking

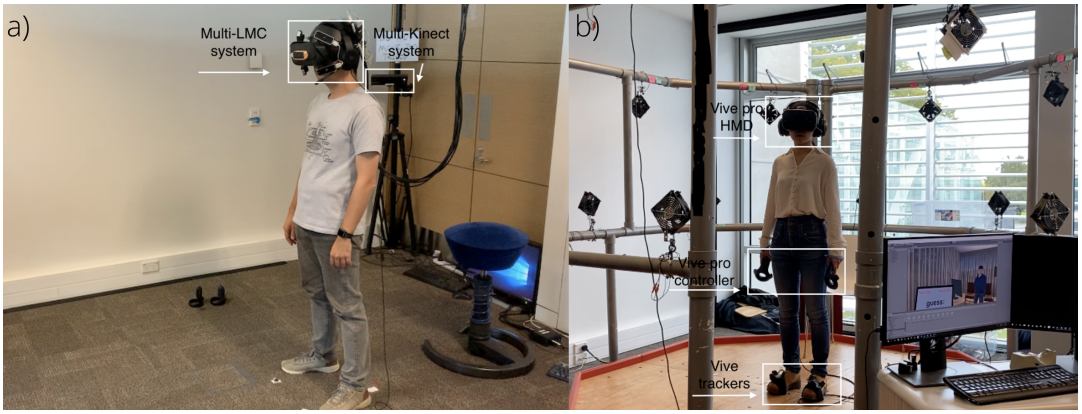


FIGURE 7.1: Asymmetric Avatar Control Systems. a) Highly expressive avatar control system, b) Low expressive avatar control system

area. This system was based on the work described in Chapters 4 and 5 with both body (21 joints including the torso, arms, and legs) and hand-gesture tracking (19 joints with pointing, grasp, and pinch). The avatar control algorithm was improved compared to Chapter 5 using a better avatar rig in the following aspects: (1) The joint rotations from the Kinects were only used as a reference, and joint rotations was re-calculated based on joint positions. (2) The skeletal tree and joint hierarchy relationship were also considered to avoid unnatural joint twists. (3) Each joint's velocity and bone direction were also calculated to make the avatar rig more smooth.

- **Hand Tracking** - the multi-LMC system was used as described in Chapter 6 with five LMCs installed on a mounting frame (Figure 7.1a).

- **Eye and Mouth Movement** - Eye-gaze and mouth movement was added to the system. The eye-gaze direction was the same as the HMD orientation, but jitter was added to simulate changes in the level of focus. The eye-gaze direction randomly changed from one object to another within a small area in the general facing direction. Also, the virtual avatar performed random eye blinking. For the mouth, fifteen visemes [73] were added to the virtual character as blend shapes. Each viseme depicted the mouth shape for a specific set of phonemes, which improved mouth-movement rendering compared to the approach described in

Chapter 5. The set of mouth shapes was driven by the *Salsa LipSync v2* [123] Unity plugin, which made the virtual mouth movement more realistic.

Low Expressive Avatar (LEA) Control System

For the low expressive avatar system, participants needed to wear tracking sensors to drive the rig of the virtual character. This was done by tracking the HMD for the head, two controllers for the hands, and two extra Vive trackers for the feet, respectively.

- **Body Tracking** - The *Final IK* [98] Unity plugin was used to calculate and estimate the positions and rotations of other joints of the body, excluding the head, hands, and feet.
- **Hand Tracking** - Virtual hand position and rotation data were mapped with the two controllers. For better human-human communication during the experiment, specific hand gestures were customized and mapped to button presses on the controllers. Squeezing the trigger button made a pointing gesture, squeezing the controller grip buttons made a "V" sign, pushing on the touchpad made a fist gesture, and doing nothing was an open hand gesture.
- **Eye and Mouth Movement** - Eye and mouth movements were similar to chapter 5, where the eye-gaze direction followed the HMD facing direction, along with random eye blinking. In contrast to the HEA, here mouth movements were approximated using small, medium, and large mouth openings, triggered by the loudness captured by the microphone using the *Salsa LipSync v1* [122] Unity plugin.

From the above description, the HEA control system is a contactless tracking medium, which can support natural social interaction, especially non-verbal behavior. The LEA control system is the general solution that is easier to setup for an avatar-based social VR application. The purpose of the user study is to compare the two systems as a whole, so the impact of a single factor

such as hands or facial expressions on communication and collaboration is not considered.

7.2.2 System Overview

Hardware and software

The HEA control system is a network solution, and the Kinect and LMC sensors are working in a client-server mode. An Intel NUC (Intel Core i5-8259U at 2.3 GHz, 8GB RAM, and Iris Plus Graphics 655) is used for the client machines to drive the connected Kinect and LMC sensors. To avoid infrared interference between the Kinects and VR devices, the inside-out tracking of the Oculus Rift S [33] was used as the HMD, which was driven by the server machine (Windows 10 desktop computer, Intel Core i7-7700K at 4.2 GHz,32GB RAM, and NVIDIA GeForce 1080Ti). The front LMC sensor was set as the primary reference connected directly to the server machine. All four client machines and the server machine were connected to a Gigabit Switch (NETGEAR GS110MX) through Ethernet cables for network data transmission (Figure 7.2a). The software was developed running on each NUC to serialize body-frame and hand-frame data retrieved from the Kinect and Leap Motion SDKs and wrapped them in Open Sound Control (OSC) messages transmitted in this local network. A standard VR setup was configured for the LEA control system (HTC Vive Pro HMD [54] with two handheld controllers). A Windows 10 desktop computer (Intel Core i7-8700 at 3.2 GHz,32GB RAM, and NVIDIA GeForce 2080) drove the Vive with two second-generation Lighthouse stations.

The SVE was developed using the Unity game engine version 2019.2.0f1 [125] with SteamVR for Unity [25], and Leap Motion plugin for Unity (Core 4.4.0) [74]. The generic virtual characters were created through Makehuman [76] with customized mouth shapes from Blender 2.79b [14]. The Point Cloud Library (PCL) [85] version 1.6.0 was used to perform LSF calculations during the multiple-LMC calibration process. The Rug.OSC library [104] was used for wrapping the transmitted frames as OSC messages and handling them in the Unity.

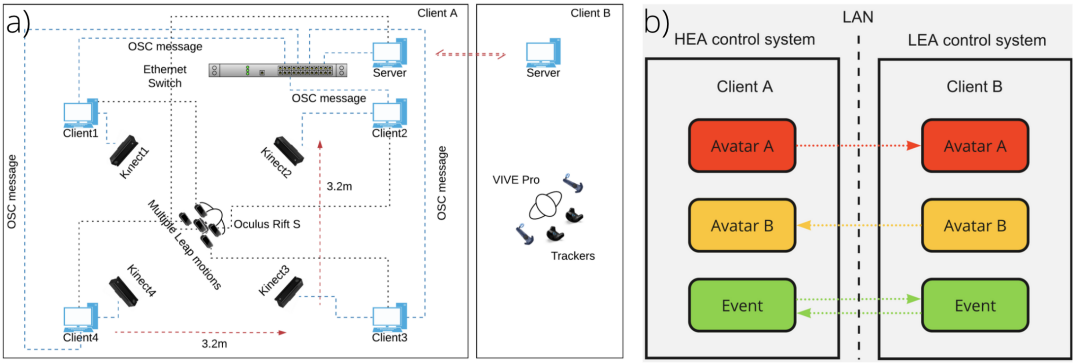


FIGURE 7.2: Multi-user VR system. a) System setup, b) Network

Networking and latency

Synchronous multi-user virtual experiences require low latency, stable connections, and accurate status synchronization. Instead of using the general server-client mode to synchronize the states for the pair of participants, a peer-to-peer (P2P) network mode was set up to directly update the avatar data and events in the SVE on both sides. Figure 7.2b shows the working mechanism. The same virtual scene was set up on both peers with either HEA or LEA configurations, and either peer could be launched first as the virtual server machine waiting for the connection. The participants who played on the Client A side (HEA control system) used the body and hand data from the multiple Kinects and LMC sensors to update the local scene first. After data fusion, Avatar A was rendered, and the system waited to transmit. Once the Unity program launched on the Client B side (LEA control system), the network connection was established. The data from the controllers and trackers drove Avatar B, which sent its data to Client A. The event listener was running on both sides and prepared for commands.

Figure 7.2a shows that the four Kinects and four of the LMCs were directly connected to the four NUCs, which continuously streamed the serialized body-frame data to the server machine at a rate of 1.5Mbps. The bandwidth requirements for each LMC was 2.4Mb/s for single-hand data and 14.4Mb/s at peak over Ethernet. All the data transmitted within the Client A system or between Client A and Client B used the UDP protocol. The OSC message handling in the HEA control system was running in the background. I changed

the message processing compared to the method in Chapter 5, so the latency of the local avatar rendering in the HEA control system was less than 10ms. The synchronized avatar data included transforms of each bone and mesh of the head (reflecting the eye and mouth movement), which can increase the latency up to 30ms. Therefore, the total latency of this multi-user VR system is less than 40ms.

The audio communication was set up using discord [56]. For an immersive sound experience, the participants wore Logitech G433 headphones on the HEA control system side and Razer Nari ultimate headphones on the LEA control system side.

7.3 Methods

A controlled laboratory experiment was conducted to investigate the impact of avatar expressiveness on communication and collaboration.

7.3.1 Participants

20 dyads, 40 participants (25 male, 15 female) were recruited from the University of Canterbury through advertisements posted on campus, and on the University social media platforms. Participants were aged 18-46 ($M = 29.3$, $SD = 6.7$), and all were students or academic staff. Basic information was collected such as level of English (13 Native speakers, 27 non-native speakers, but all could speak English fluently) and dyad relationships (34 friends, 6 classmates or colleagues). Participants were asked about their familiarity with VR using a 5-point Likert scale, from 1 (never), 3 (a few times a month), to 5 (daily use). The participants generally had moderate experience using VR ($M = 2.4$, $SD = 0.93$). The frequency of social VR platform use was never (62.5%), a few times a year (32.5%), and a few times a month (5%). From the demographic information, most participants had VR experience, but only 37.5% of subjects had tried social VR applications before. As I used a charades

game, all the participants were asked about their charades expertise, never (37.5%), beginner (37.5%), intermediate (22.5%), and expert (2.5%).

7.3.2 Study Design

The present study adopted a within-subjects design with one independent variable (expressiveness) with two levels: highly-expressive avatar (HEA) and low-expressive avatar (LEA) as described previously. To evaluate the user behavior and experience in different avatar control systems during mutual communication and collaboration, a charades game playing scenario was set up. The experiment had four game-play sessions per dyad. In each session, the dyad used both sides and embodied the relevant avatar, either the word performer or the guesser. The purpose was to make sure the dyad could try both avatar systems and take turns in the different roles. The participants were asked to rate their experience with the system after using one system in both roles (word performer, guesser). Participants' orders were randomized using Research Randomizer [131].

The Scene and Charades Game

The SVE was a virtual living room, with the two virtual characters placed facing each other as shown in Figure 7.3. The distance between the players was around 2 meters. A virtual display was placed on a small table in front and to the side of each avatar to show the words to mime, and the number of words left. There was a countdown timer displayed on the wall once the game started. In the physical world, an experimenter sat on the Client A side and used a keyboard to control the whole process. After the pair of participants put on the HMDs and were ready for the study, the experimenter pressed a button to start the game, and the participants in the virtual world could see a text message about the game start from a first-person perspective. The experiment consisted of four sessions. For each session, a set of ten words was selected from [57, 39] with different difficulty (six easy words and four hard words). The sets were:

- **Set 1:** pillow, tail, drum, mouth, finger, hungry, haircut, password, fast food, traffic jam.
- **Set 2:** swimming, love, hugs, itchy, grab, basketball, glue gun, sushi, cushion, police.
- **Set 3:** boxing, weightlifting, lobster, applaud, dancing, walking, lunch box, painting, elevator, earthquake.
- **Set 4:** scissor, crouching, hammer, piano, guitar, robot, thief, assemble, barber, pocket.

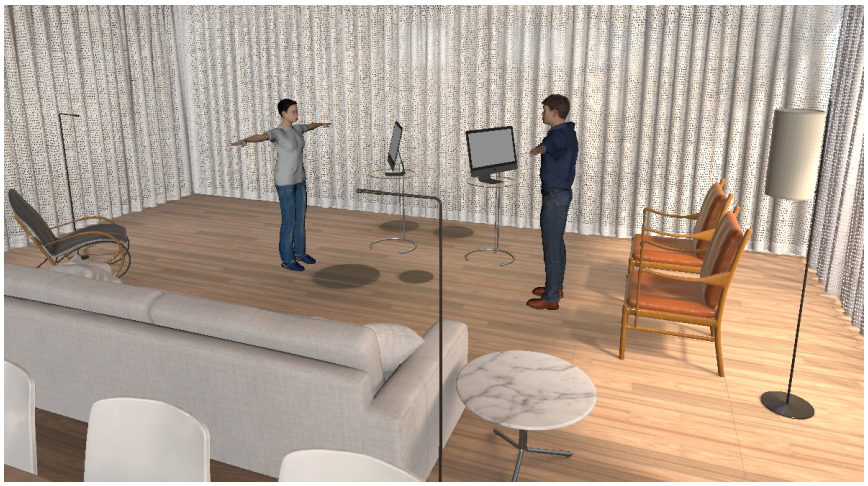


FIGURE 7.3: The charade game scene

Once the game started, the participant at Client A saw a word shown on the virtual display, and he/she could only use non-verbal cues such as body posture and hand gestures to describe the word. The other participant could use verbal and non-verbal cues to guess or ask the performer for more. They needed to collaborate to finish the ten words within five minutes. In the second session, the participants stayed in their positions but switched roles. The player at Client B mimed the next set of words for the player at Client A. For the next two sessions, the participants swapped the avatar systems and repeated the first two sessions with different word sets.

Figure 7.4 shows the four sessions, and the virtual view in each picture is from the partner. During the game, the experimenter listened to the guesser,

and if he/she said the correct word, the experimenter pressed the "Next" button, and in the virtual world, the participants could jump to the next word. If the participants thought the current word was too hard to perform or his/her partner was taking too long to guess it, the guesser could ask the experimenter to pass the word. The experimenter would then press the "Pass" button to skip the word, and the system would record which word was passed for later analysis.

Hypotheses

I expected that more-realistic avatar control would make participants perform more naturally and feel more socially connected during the game. Based on this expectation, and the previous related work in the field, several hypotheses were formulated.

- H_1: Participants will feel greater copresence interacting with the highly expressive avatar in the collaborative environment.
- H_2: Participants will feel greater social presence interacting with the highly expressive avatar in the collaborative environment.
- H_3: Participants will feel greater attraction interacting with the highly expressive avatar in the collaborative environment.
- H_4: Participants using the highly expressive avatar control system will perform better on the communicative task than those who use a low expressive avatar control system.

7.3.3 Measurements

Data were collected in two ways. Most subjective questionnaires were filled out after every two sessions. The system automatically recorded the completion time and number of passed words.

FIGURE 7.4: The experimental process. a) Session 1, word performer using HEA control, b) Session 2, word performer using LEA control, c) Session 3, word performer using HEA control, d) Session 4, word performer using LEA control

Copresence

Copresence is the feeling that the user is with other entities [107]. The copresence in this user study was measured by two separate scales, their involvement in the interaction (self-reported copresence) and perception of their partner's involvement in the interaction (perceived other's copresence). The questionnaires for copresence were from Nowak et al. [90], which were also used in the previous research from Roth et al. [102]. The self-reported copresence scale included six items asking the participants to self-report their level of involvement in the interaction. The perceived other's presence scale included twelve indicators for intimacy, involvement, and immediacy. Participants rate their level of agreement with statements like, "I was interested in talking to my interaction partner" and "The interaction partner communicated coldness rather than warmth", on a 7-point Likert scale (1 = strongly agree, 7 = strongly disagree). The reliability of the scales were tested using the data collected

in the experiment, and found the copresence scales had good internal consistency: self-reported presence (Cronbach's α = 0.726), perceived other's presence (Cronbach's α = 0.810).

Social presence

Social presence is the feeling of the user, which makes people feel connected with others through the telecommunication system, according to Rice [97], Short et al., [110] and Walther [135]. The questionnaire for social presence was from Nowak et al. [90]. The scale consisted of six items, and participants used a sliding scale (0-100) to answer questions like "To what extent did was this like you were in the same room with your partner?". The reliability of the scale was good (Cronbach's α = 0.768).

Interpersonal attraction

The measure for liking and attraction was adapted from Oh et al. [91], which consisted of six items. Sample items include "I would enjoy a casual conversation with my partner" and "I would get along well with my partner". It was using a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree). The reliability of the scale was good (Cronbach's α = 0.921).

Finally, the participants were asked to fill out the post-questionnaire about system preference and comments. Sample items include: "Which VR system was most helpful when you were describing words to your partner?" and "Which VR system do you prefer?"

7.3.4 Procedure

The participants were asked to fill out the demographic survey and consent form before beginning the experiment. Then, the experimenter introduced the rules and the whole process and explained how to use the devices involved in this user study. Charades is a communicative and collaborative game that requires players to use specific body postures or hand gestures. The rules for describing the words, and the level of expertise, can vary from person

to person, so both participants were required to discuss strategies before the experiment began. Therefore, the researcher arranged a face-to-face discussion before the game to reduce the risk of a bad game experience with different opinions.

After the discussion, both participants were guided to their respective avatar control systems. The experimenter helped them put on the HMD, gave them the relevant devices, and let them get familiarized with the system. Once the connection was established, the participants on both sides were asked to practice communication only using non-verbal behavior. Then the Discord program was launched for an audio communication test.

When they were ready for the game, the experimenter started the game for the first two sessions. After that, the participants were required to fill out questionnaires. The experimenter then cleaned all of the devices and changed the configuration so that the participants could swap avatar control systems for the remaining sessions. Finally, participants were given one additional survey to gather information about their preference and ease of use of the avatar control schemes. The researcher then performed an experimental debrief with the participants, encouraged them to write comments about the two systems, discuss their survey answers, and talk about their general impressions of the two systems.

7.3.5 Statistical Analysis

For the analysis, the collected data sets of 40 participants (20 dyads) were used. A paired-samples t-test was used to compare participants' ratings of copresence, social presence, and interpersonal attraction for the two system. In Table 7.1, t-test values, means and standard deviations for the questionnaires are presented. α = 0.05 was used as level for statistical significance. Shapiro-Wilk test was implemented before the t-test to check if the collected data is normally distributed and found that self-reported copresence (HEA($p = 0.203$), LEA($p = 0.054$)), perceived partner's copresence (HEA($p = 0.875$), LEA($p = 0.069$)), social presence (HEA($p = 0.064$),

LEA($p = 0.395$)), interpersonal attraction (HEA($p = 0.432$), LEA($p = 0.056$)) did not significantly deviate from it. The Shapiro-Wilk test for users performance data on completion time (Group A ($p = 0.916$), Group B ($p = 0.119$)) and Number of passed words (Group A ($p = 0.977$), Group B ($p = 0.817$)) was also not significant.

7.4 Results

In this section, the summarized data and results of the statistical analyses is provided. Table 7.1 and Table 7.2 as well as Figure 7.5 and Figure 7.6 provide overview of the collected data. The questionnaires to measure the social presence and interpersonal attraction were used from [90, 91] focus on the experience by reviewing a partner's performance. Hence the scores in the table are based on the system that their counterpart used.

TABLE 7.1: Statistical results for copresence, social presence, and interpersonal attraction

	Copresence		**Social presence**	**Interpersonal attraction**
	Self-reported	Preceived other's		
t-test	$p = 0.661$	$p = 0.819$	$p = 0.0008$	$p = 0.0007$
HEA M (SD)	4.2 (0.60)	4.0 (0.47)	63.0 (18.87)	5.4 (1.23)
LEA M (SD)	4.2 (0.59)	4.0 (0.46)	72.8 (7.99)	6.1 (0.46)

7.4.1 Social presence

There was a significant difference ($t(39) = 3.632, p < 0.001$) on how participants rated social presence for the two systems. Participants interacting with a HEA counterpart rated social presence significantly higher ($M = 72.8, SD = 7.99$) than when they interacted with a LEA counterpart ($M = 63.0, SD = 18.87$).

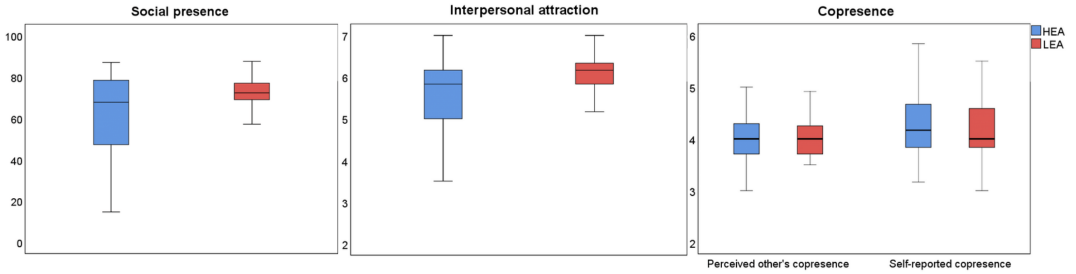


FIGURE 7.5: Statistical results. a) Social presence, b) Interpersonal attraction, c) Copresence

7.4.2 Interpersonal attraction

Similarly, participants ratings showed that there was a significant difference for interpersonal attraction ($t(39) = 3.685, p < 0.001$) again showing higher results for participants interacting with a HEA counterpart ($M = 6.1, SD = 0.46$) compared to the LEA condition ($M = 5.4, SD = 1.23$).

7.4.3 Copresence

The collected data did not show any significant differences between the HEA and LEA systems for copresence, neither the sub-component self-reported copresence ($t(39) = 0.442, p = 0.661$) nor the perceived partner's copresence ($t(39) = 0.231, p = 0.819$).

TABLE 7.2: Summary of objective measurement results

Session	**Time to complete**	**Number of passed words**	**Session role**	
			HEA	LEA
1 and 3 M (SD)	290.3 (8.5)	1.8 (0.8)	performer	guesser
2 and 4 M (SD)	291.5 (7.4)	3.1 (1.4)	guesser	performer

7.4.4 Performance

The completion time and the number of passed words were recorded. For each session, participants saw a timer of five minutes to finish displayed in the virtual world, but they were allowed to continue if they did not manage to

go through all ten words within that time. The collected data were splited into two groups. Group A for conditions in which participants were using HEA as the performer and partners using LEA as the word guesser and Group B in which participants were using LEA as the performer and partners using HEA as the word guesser. There was no significant difference between the amount of time participants took to finish each session when the performer used either HEA ($M = 290.3, SD = 8.5$) or LEA ($M = 291.5, SD = 7.4$) to describe the words ($t(39) = 0.698, p = 0.489$). The results, however, show that there is a significant difference ($t(39) = 5.551, p < 0.001$) between the two groups for the number of passed words. When participants used the HEA control system to describe the words, they passed fewer words ($M = 1.8, SD = 0.8$) compared to performers using the LEA ($M = 3.1, SD = 1.4$).

7.4.5 Preference

The results show that 31 (77.5%) of participants thought it was easier to use HEA as a word performer to describe words to their partner. Furthermore, about 35 (87.5%) of participants preferred the HEA overall.

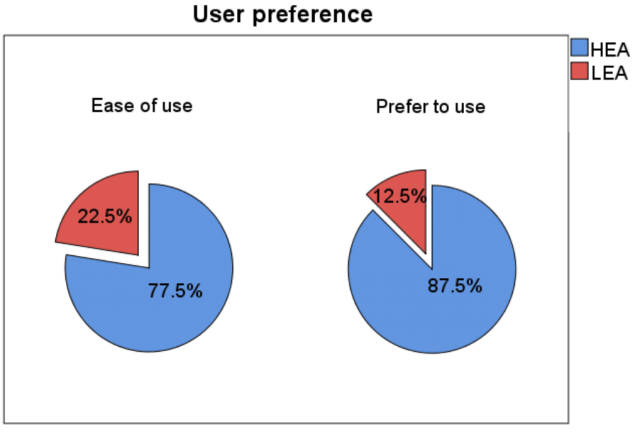


FIGURE 7.6: Preference

Participants also provided comments about the experiment and their embodied avatar control experience during mutual collaboration. Many comments reflect the importance of natural and accurate non-verbal behavior for high-quality communication, which can let user immerse themselves in the SVE and experience communication more like a face-to-face meeting.

- *"High quality of experience about the person-to-person meeting, easy to understand what my partner wants to show/say. To compare, the HEA control system brings more real experience. it shows a clear movement of my partner's whole body."*

- *"In the LEA control system, I sometimes felt despair because I knew a simple gesture that would have explained the word immediately, but I could not do it and could not come up with something to replace or mimic it with the limited capabilities.*

- *"I tried both systems, I prefer the HEA rather than LEA. The HEA control system is like the real world much more than the LEA control system, more activities, more details. It feels I have more communication between us. Besides, when I used the controller, I only can use my arms, legs and two fingers.*

- *"I like the HEA control system because it is flexible. It was still quite different from real-life face-to-face experience, but It acts as a benefit to me like I don't feel shy to perform something that I might not perform in real life.*

7.5 Discussion

In this study, it was found that participants who interacted with people using avatars that had highly expressiveness, non-verbal behavior felt greater social presence, which supports hypothesis H_2. Furthermore, the participants felt more attracted when they communicated and collaborated with the users who used the HEA control system, which supports hypothesis H_3. Another important aspect to note is that the majority of the participants preferred the HEA control system and felt it was easier to use. As for hypothesis H_4, the statistical results partly support it. The average number of successfully explained words for a user using HEA as a performer was 8.2 (82%), which is higher than the condition when participants used the LEA control system as a performer 6.9 (69%). However, there was no statistical difference between completion times. This could partially because participants were presenting a

timer of 5 minutes, which led to a ceiling effect that most participants took close to 5 minutes. As other factors, such as the amount of skipped word, can also impact the completion time for hypothesis H_4. Hence, the number of completed words can be be seen as the only suitable measure of users' performance.

I noticed that the participants have different behavior during the experiment. The participants who were using the HEA control system would like to move more such as complicated body posture, and hand gestures. Most of them feel confident as they can perform as they intend to do. On the contrary, the users who were using the LEA control system have less confidence and move less. The tracking performance of the HEA system restricted the performance of participants.

I did not find evidence to support hypothesis H_1. The embodied experience can provide a similar sense of presence when the participants use a simple avatar control system, as was found in Chapter 5. Therefore, if the SVE system is stable with low network latency, and the participants can both communicate with each other based on their real behavior, it is not hard to understand that there is no significant difference between the high and low expressive avatar control system in the either self-reported and perceived copresence. However, from the user comments and the pie chart in Figure 7.6, it can be concluded that participants preferred using the HEA control system to communicate and collaborate in VR because it was flexible and more natural.

7.5.1 Implications

The findings have practical implications for designers and developers of shared virtual environments. A highly expressive avatar control system that can support natural non-verbal behavior can lead to a more positive and realistic experience between players. It is intuitive and straightforward to express themselves with body posture or hand gestures when they communicate and collaborate in the SVE. The positive effects on social presence and

interpersonal attraction from the highly expressive avatar control system can make virtual communication more like a face-to-face experience.

7.5.2 Limitations

Some limitations of the study need to be addressed. First, some participants reported that the HMD was a little bit heavy for the HEA control system due to the presence of five LMCs mounted on the HMD, along with the necessary extension cables. Although I tried managed the cables by hanging them from the ceiling, they still may have bothered participants during gameplay. Second, some participant actions went beyond the hand tracking area, even though the system greatly enlarges the area compared to normal tracking. For example, sometimes they moved their hands over their heads. Also, participants sometimes touched the mounting frame of the multi-LMC system on the HMD, which in some cases resulted in the need to re-calibrate the system to guarantee quality hand tracking. Therefore, the participants were asked to avoid touching the sensors on the HMD and reduce arm movement amplitude when they moved their hands over their heads. This could have affected the participant's perceptual and cognitive load. Third, in this study, the participants were paired regardless of gender. The performance may be different when females collaborate with males compared to other gender combinations. The gender needs to be considered as a factor when I design collaborative studies.

7.6 Conclusions and Future Work

A shared virtual environment was implemented using an asymmetric avatar control system and investigated the impact of different levels of non-verbal expressiveness on communication and collaboration behavior through a virtual charades game. I found a significantly higher social presence and interpersonal attraction when participants interacted with a user using the HEA

control system. Participants prefer using the highly expressive avatar control system, which improves the task performance with a higher number of successful explained words.

In future work, I plan to improve the multi-LMC system by replacing the five extension cables with wireless transmitters and receivers. I also plan to refine the calibration algorithm for the multi-LMC system to a self-adaptive version, so the player does not need to re-calibrate the system if the frame mount is moved. Furthermore, I consider to add tactile feedback into this multi-user VR system to explore the effect of haptic cues on communication and collaboration behavior.

Chapter 8

Conclusions and Future Work

This thesis has investigated the impact of avatar expressiveness, especially non-verbal behavior, on single-user and multi-user VR experiences. To conduct this research, a highly expressive avatar control system was designed and built with robust full body and hand gesture tracking, along with realistic eye and facial movement rendering. Then two types of user studies were designed and implemented. One was used to compare sensor-based and controller-based avatar control systems in a single-user simulated communication application. The other one compared different levels of expressiveness in avatars for collaborative applications using asymmetric avatar control systems.

Relevant research about the effects of non-verbal behavior realism in avatar-mediated communication and collaboration was reviewed in Chapter 2. In the later subsections, avatar-related VR properties such as presence, body ownership, and embodiment were provided. The next part of Chapter 2 was an evaluation of a literature review of avatar control systems in terms of body and hand tracking, data fusion, and avatar rendering in the subsequent subsection. The last part of the chapter covered the tracking systems and methods.

In Chapter 3, a multi-Kinect system was designed to provide fully articulated body tracking regardless of the user's orientation. The goal was to present a system that can provide a contactless full-body tracking experience with robust tracking. This laid the foundation for later avatar systems, as the body tracking data is the primary data source.

In Chapter 4, an improved avatar tracking system was presented based on the work in Chapter 3. The number of Kinect sensors was increased to four for the larger tracking area, and an additional calibration procedure eliminated the offset error between depth and RGB cameras. The facing direction calculation was improved for robust data fusion, and the proposed adaptive weight calculation method improved data fusion accuracy. With the integration of the Leap Motion controller, the avatar system presented in this chapter provides a fully tracked body and hands experience in VR through a first-person perspective.

In Chapter 5, a study was presented in VR where the effect of a depth-sensor based avatar control system on user communication behavior was investigated. In the study, the depth sensor-based avatar control system with a controller-based avatar control system was compared through a virtual interview. The participants went through two interview sessions using both systems. The virtual interview process in VR was recorded, and the participants evaluated their performance through a third-person view. It was found that significantly higher virtual body ownership illusion and usability, as well as better non-verbal communication performance, by participants in the depth-sensor-based experience compared to the controller-based experience.

In Chapter 6, a multi-LMC system was designed to enlarge the usable hand tracking area for my avatar control system. A novel shared-view calibration method was proposed based on the LSF algorithm and a multi-LMC fusion algorithm based on two-level data evaluation (prediction-based and position-based evaluation methods). The system experiment showed that the system can enlarge the hand tracking range to 202.16 degrees horizontally and 164.43degrees vertically.

In Chapter 7, the multi-LMC system was integrated with the avatar control system from Chapter 5. This optimized the avatar control algorithm, and improved the mouth and eye movement rendering. The effects of different levels of avatar expressiveness were investigated on communication and collaboration behavior through a virtual charades game. The participants took turns as word performers and word guessers using the different avatar control

systems, and needed to collaborate to complete four sessions within the given time. It was found that a significantly higher social presence and interpersonal attraction was achieved when the participants interacted with the user who was using the highly expressive avatar control system. Additionally, participants had better task performance when they embodied a highly expressive avatar.

8.1 Contribution

The main contribution of this thesis is the development and evaluation of a novel highly expressive avatar control system, specifically for supporting natural non-verbal behavior for communication and collaboration in a shared virtual environment. Evidence was provided that natural and intuitive non-verbal behavior such as body posture, hand gestures, and even facial expressions can augment social interaction in virtual environments. The implications of this may benefit many domains in which embodied avatar communication scenarios are used.

This thesis describes how to integrate off-the-shelf tracking technology into an avatar system that provides natural and accurate full-body movement and hand gesture tracking as well as realistic facial expressions. Specifically, the user does not need to wear any tracking sensors. This is an under-explored research domain, which makes this research a valuable contribution to this space.

The expressive sensor-based avatar control system demonstrated that it has improved usability, body ownership illusion, and sense of agency compared to a controller-based avatar system. Furthermore, users also reported a strong preference for the expressive avatar system that does not require additional body-worn sensors. It was shown that the highly expressive avatar can enhance the social presence and performance in a collaborative task. The greater interpersonal attraction that the user can obtain when interacting with a user using a highly expressive avatar control system makes mutual communication more like in a face-to-face scenario. In summary, the results suggest

that social VR applications can benefit from the use of highly expressive avatars that extend non-verbal communication possibilities.

8.2 Limitations

There were some limitations in the system designs and participant recruitment.

Multi-Kinect calibration: In Chapter 3, multiple Kinects were calibrated through the chessboard marker captured by the RGB camera, but the skeleton data was from a depth camera, which included some offset error when calibrating the coordinate system from Kinect to the chessboard. Although the error was eliminated through an additional calibration method in Chapter 4, it would be better to calibrate directly through the depth camera.

Multi-LMC system: The HMD was a little bit heavy due to the presence of five LMCs mounted on the HMD, along with the necessary extension cables. Although efforts were made to manage the cables by hanging them from the ceiling, they still may have bothered participants during gameplay. Additionally, some participant actions went beyond the hand tracking area, even though the system greatly enlarges the area compared to normal tracking. For example, sometimes, users' hands were moved of their heads. Also, participants sometimes bumped into the mounting frame of the multi-LMC system, and re-calibration was required for the system to guarantee quality hand tracking. This could have affected the participant's perceptual and cognitive load.

Participants: All the studies were conducted at the University of Canterbury. Participants were recruited on campus, and some of the subjects participated in more than one study. The participants were paired regardless of gender in Chapter 7. The performance may be different when females collaborate with males compared to other gender combinations. Thinking about gender when designing a collaborative study is important.

Avatar control system: In Chapter 7, the movement of eye and mouth rendering was more realistic compared to Chapter 5. However, data was not

captured by external devices (e.g., a camera), which could more realistically reflect real-time captured data.

8.3 Future Work

Lessons were learned from the avatar system and the implemented studies. There are some directions for future work to continue the research undertaken in this thesis.

The effects of emotional cues were not investigated, such as facial expressions, which are part of avatar expressiveness, and how they affect real-time collaboration. The work from Roth et al. [100, 99] showed that the captured facial expression could support self-disclosure and augment social behavior. Therefore, in the future, exploring the effects of emotion cues on the communication and collaboration behavior in VEs would be beneficial.

Tactile feedback is another channel of avatar expressiveness. The consistency of haptic sensations in the physical and virtual world can enhance the presence and immersion in VEs. In the future, adding tactile feedback into this multi-user VR system would help to explore the effects of capturing, transmitting and displaying haptic cues on communication and collaboration behavior.

In Chapters 3, 4, and 6, the avatar system was presented with improved features and functionality. After optimizing and improving the avatar control system, there were still limitations. In the future, upgrading the system in the following aspects would be beneficial: 1) For the multi-LMC system, replace the five extension cables with wireless transmitters and receivers, and refine the calibration algorithm for a self-adaptive version. 2) Optimizing the hands' frame transmission. 3) Integrating the facial expression tracking device and haptic sensors to the avatar system.

8.4 Conclusions

A highly expressive avatar control system was built that supports robust and natural non-verbal behavior, which answered the research questions 2, 3 and 5. Two user studies were designed and implemented to validate the system and explored the system's effect on communication and collaboration behavior in the single user and multiple users applications for research questions 4 and 6. The results showed that the system could support communication with better usability, a great sense of body ownership, and agency. The users exhibited better task performance in the SVE with the collaborative task using the highly expressive avatar system. The users presented greater social presence and interpersonal attraction when interacting with the users using a highly expressive avatar control system. Hence, it can be concluded that virtual reality avatar systems benefit from a higher level of non-verbal expressiveness, which can be achieved without additional body-worn trackers.

Bibliography

[1] Mohd Hezri Amir, Albert Quek, Nur Rasyid Bin Sulaiman, and John See. "DUKE: Enhancing Virtual Reality based FPS Game with Full-body Interactions". In: *Proceedings of the 13th International Conference on Advances in Computer Entertainment Technology - ACE2016* (2016), pp. 1–6. DOI: 10.1145/3001773.3001804.

[2] Nadeem Anjum and Andrea Cavallaro. "Trajectory Association and Fusion across Partially Overlapping Cameras". In: *Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2009, 2-4 September 2009, Genova, Italy*. Genova, Italy: IEEE, 2009, pp. 201–206. DOI: 10.1109/AVSS.2009.65.

[3] Andreas Aristidou, Joan Lasenby, Yiorgos Chrysanthou, and Ariel Shamir. "Inverse Kinematics Techniques in Computer Graphics: A Survey". In: *Computer Graphics Forum*. Vol. 37. 6. Wiley Online Library. 2018, pp. 35–58.

[4] K. S. ARUN. "Least-squares fitting of two 3-D point sets". In: *IEEE Trans Pattern Anal Mach Intell* 9 (1987), pp. 698–700. DOI: 10.1109/TPAMI.1987.4767965.

[5] Jeremy N Bailenson, Andrew C Beall, Jack Loomis, Jim Blascovich, and Matthew Turk. "Transformed social interaction: Decoupling representation from behavior and form in collaborative virtual environments". In: *Presence: Teleoperators & Virtual Environments* 13.4 (2004), pp. 428–441.

[6] Jeremy N. Bailenson, Andrew C. Beall, and Jim Blascovich. "Gaze and task performance in shared virtual environments". In: *The Journal of Visualization and Computer Animation* 13.5 (2002), pp. 313–320. DOI: 10.

1002/vis.297. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/vis.297. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/vis.297.

[7] Barbara Becker and Gloria Mark. "Social Conventions in Computermediated Communication: A Comparison of Three Online Shared Virtual Environments". In: *The Social Life of Avatars: Presence and Interaction in Shared Virtual Environments*. Ed. by Ralph Schroeder. London: Springer London, 2002, pp. 19–39. ISBN: 978-1-4471-0277-9. DOI: 10.1007/978-1-4471-0277-9_2. URL: https://doi.org/10.1007/978-1-4471-0277-9_2.

[8] Giovanni Berlucchi and Salvatore Aglioti. "The body in the brain: neural bases of corporeal awareness". In: *Trends in neurosciences* 20.12 (1997), pp. 560–564.

[9] Frank Biocca. "The cyborg's dilemma: Progressive embodiment in virtual environments". In: *Journal of computer-mediated communication* 3.2 (1997), JCMC324.

[10] Gary Bishop, Greg Welch, et al. "An introduction to the kalman filter". In: *Proc of SIGGRAPH, Course* 8.27599-23175 (2001), p. 41.

[11] Olaf Blanke. "Multisensory brain mechanisms of bodily self-consciousness". In: *Nature Reviews Neuroscience* 13.8 (2012), pp. 556–571.

[12] Olaf Blanke and Thomas Metzinger. "Full-body illusions and minimal phenomenal selfhood". In: *Trends in cognitive sciences* 13.1 (2009), pp. 7–13.

[13] Jim Blascovich. "Social influence within immersive virtual environments". In: *The social life of avatars*. Springer, 2002, pp. 127–145.

[14] blender. *blender.org - Home of the Blender project - Free and Open 3D Creation Software*. 2019. URL: https://www.blender.org/ (visited on 06/25/2019).

[15] Dario Bombari, Marianne Schmid Mast, Elena Canadas, and Manuel Bachmann. "Studying social interactions through immersive virtual environment technology: virtues, pitfalls, and future challenges". In: *Frontiers in Psychology* 6.June (2015), pp. 1–11. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2015.00869. URL: http://journal.frontiersin.org/Article/10.3389/fpsyg.2015.00869/abstract.

[16] Dario Bombari, Marianne Schmid Mast, Elena Canadas, and Manuel Bachmann. "Studying social interactions through immersive virtual environment technology: virtues, pitfalls, and future challenges". In: *Frontiers in psychology* 6 (2015), p. 869.

[17] Matthew Botvinick and Jonathan Cohen. "Rubber hands 'feel' touch that eyes see". In: *Nature* 391.6669 (1998), p. 756.

[18] Edward Brent and G Alan Thompson. "Sociology: Modeling social interaction with autonomous agents". In: *Social Science Computer Review* 17.3 (1999), pp. 313–322.

[19] John Brooke et al. "SUS-A quick and dirty usability scale". In: *Usability evaluation in industry* 189.194 (1996), pp. 4–7.

[20] Joseph N Cappella. "Mutual influence in expressive behavior: Adult–adult and infant–adult dyadic interaction." In: *Psychological bulletin* 89.1 (1981), p. 101.

[21] Polona Caserman, Augusto Garcia-Agundez, Robert Konrad, Stefan Göbel, and Ralf Steinmetz. "Real-time body tracking in virtual reality using a Vive tracker". In: *Virtual Reality* 0123456789 (2018). ISSN: 1359-4338. DOI: 10.1007/s10055-018-0374-z. URL: http://link.springer.com/10.1007/s10055-018-0374-z.

[22] Simon Chapple and Maxime Ladaique. *Society at a Glance 2009: OECD social indicators*. Paris,France: Organisation for Economic Co-operation and Development, 2009.

[23] ALEX COLGAN. *How Does the Leap Motion Controller Work?* 2014. URL: http://blog.leapmotion.com/hardware-to-software-how-does-the-leap-motion-controller-work/ (visited on 06/25/2019).

[24] Tara Collingwoode-Williams, Marco Gillies, Cade McCall, and Xueni Pan. "The effect of lip and arm synchronization on embodiment: A pilot study". In: *Proceedings - IEEE Virtual Reality* (2017), pp. 253–254. ISSN: 0190-8286. DOI: 10.1109/VR.2017.7892272.

[25] Valve Corporation. *steam vr*. 2019. URL: https://www.steamvr.com/en/ (visited on 05/15/2020).

[26] Adam Craig and Sreenath Krishnan. *Fusion of Leap Motion and Kinect Sensors for Improved Field of View and Accuracy for VR Applications*. 2016.

[27] Katarzyna Czesak, Raul Mohedano, Pablo Carballeira, Julian Cabrera, and Narciso Garcia. "Fusion of pose and head tracking data for immersive mixed-reality application development". In: (2016), pp. 1–4.

[28] Laura Dipietro, Angeloi M. Sabatini, and Paolo Dario. "A survey of glove-based systems and their applications". In: *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews* 38.4 (2008), pp. 461–482. ISSN: 10946977. DOI: 10.1109/TSMCC.2008.923862.

[29] Trevor J Dodds, Betty J Mohler, and Heinrich H Bülthoff. "Talk to the virtual hands: Self-animated avatars improve communication in head-mounted display virtual environments". In: *PloS one* 6.10 (2011), e25759.

[30] D Christopher Dryer. "Getting personal with computers: how to design personalities for agents". In: *Applied artificial intelligence* 13.3 (1999), pp. 273–295.

[31] Goffman Erving. "Behavior in public places: notes on the social organization of gatherings". In: *New York* (1963).

[32] M Fabri, DJ Moore, and DJ Hobbs. "Expressive agents: Non-verbal communication in collaborative virtual environments". In: *Proceedings of Autonomous Agents and Multi-Agent Systems (Embodied Conversational Agents)* (2002).

[33] Facebook. *Oculus Rift S*. 2019. URL: https://www.oculus.com/rift-s/ (visited on 06/25/2019).

[34] Tiare Feuchtner and Jörg Müller. "Ownershift: Facilitating Overhead Interaction in Virtual Reality with an Ownership-Preserving Hand Space Shift". In: *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 2018, pp. 31–43.

[35] Mary Ellen Foster. "Enhancing human-computer interaction with embodied conversational agents". In: *International Conference on Universal Access in Human-Computer Interaction*. Springer. 2007, pp. 828–837.

[36] J. Fountain and S. P. Smith. "Real-time Ambient Fusion of Commodity Tracking Systems for Virtual Reality". In: *Proceedings of the 27th International Conference on Artificial Reality and Telexistence and 22Nd Eurographics Symposium on Virtual Environments*. ICAT-EGVE '17. Adelaide, Australia: Eurographics Association, 2017, pp. 1–8. URL: http://dl.acm.org/citation.cfm?id=3298830.3298832.

[37] Sebastian Friston and Anthony Steed. "Measuring latency in virtual environments". In: *IEEE transactions on visualization and computer graphics* 20.4 (2014), pp. 616–625.

[38] Woodrow Barfield Thomas A Furness. *Virtual environments and advanced interface design*. Oxford University Press on Demand, 1995.

[39] The Game Gal. *Game Word Generator - The Game Gal*. 2020. URL: https://www.thegamegal.com/word-generator/ (visited on 04/16/2020).

[40] Shaun Gallagher. "Philosophical conceptions of the self: implications for cognitive science". In: *Trends in cognitive sciences* 4.1 (2000), pp. 14–21.

[41] Maia Garau, Mel Slater, Vinoba Vinayagamoorthy, Andrea Brogni, Anthony Steed, and M Angela Sasse. "The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2003, pp. 529–536.

[42] Maia Garau, Mel Slater, Vinoba Vinayagamoorthy, Andrea Brogni, Anthony Steed, and M. Angela Sasse. "The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment". In: *Proceedings of the conference on Human factors in computing systems - CHI '03* 5 (2003), p. 529. DOI: 10.1145/642700.642703. URL: http://portal.acm.org/citation.cfm?doid=642611.642703.

[43] Darren Gergle, Robert E Kraut, and Susan R Fussell. "Using visual information for grounding and awareness in collaborative tasks". In: *Human–Computer Interaction* 28.1 (2013), pp. 1–39.

[44] Richard J Gerrig. *Experiencing narrative worlds: On the psychological activities of reading*. Yale University Press, 1993.

[45] Marco Gillies and Mel Slater. "Non-verbal communication for correlational characters". In: (2005).

[46] Mar Gonzalez-Franco, Daniel Perez-Marcos, Bernhard Spanlang, and Mel Slater. "The contribution of real-time mirror reflections of motor actions on virtual body ownership in an immersive virtual environment". In: *2010 IEEE virtual reality conference (VR)*. IEEE. Waltham, MA, USA: IEEE, 2010, pp. 111–114.

[47] Mar Gonzalez-Franco Gonzalez-Franco and Tabitha C Peck. "Avatar Embodiment. Towards a Standardized Questionnaire." In: *Frontiers in Robotics and AI* 5 (2018), p. 74.

[48] Michael SA Graziano. "How the brain represents the body: insights from neurophysiology and psychology". In: *Common mechanisms in perception and action* (2000).

[49] Jože Guna, Grega Jakus, Matevž Pogačnik, Sašo Tomažič, and Jaka Sodnik. "An analysis of the precision and reliability of the leap motion sensor and its suitability for static and dynamic tracking". In: *Sensors* 14.2 (2014), pp. 3702–3720. DOI: 10.3390/s140203702. URL: https://doi.org/10.3390/s140203702.

[50] Sandra G. Hart and Lowell E. Staveland. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research". In: *Human Mental Workload*. Ed. by Peter A. Hancock and Najmedin Meshkati. Vol. 52. Advances in Psychology. North-Holland, 1988, pp. 139–183. DOI: https://doi.org/10.1016/S0166-4115(08)62386-9. URL: http://www.sciencedirect.com/science/article/pii/S0166411508623869.

[51] Carrie Heeter. "Being there: The subjective experience of presence". In: *Presence: Teleoperators & Virtual Environments* 1.2 (1992), pp. 262–271.

[52] Paul Heidicker, Eike Langbehn, and Frank Steinicke. "Influence of avatar appearance on presence in social VR". In: *2017 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE. 2017, pp. 233–234.

[53] Alexander Hornung, Sandip Sar-Dessai, and Leif Kobbelt. "Self-calibrating optical motion tracking for articulated bodies". In: *IEEE Proceedings. VR 2005. Virtual Reality, 2005.* 2005 (2005), pp. 75–82. ISSN: 1087-8270. DOI: 10.1109/VR.2005.1492756.

[54] HTC. *VIVE pro | The professional-grade VR headset*. 2018. URL: https://www.vive.com/nz/product/vive-pro/ (visited on 04/15/2020).

[55] TianJian Hu, XiaoJun Zhu, XueQian Wang, TianShu Wang, JunFeng Li, and WeiPing Qian. "Human stochastic closed-loop behavior for master-slave teleoperation using multi-leap-motion sensor". In: *Science China Technological Sciences* 60.3 (2017), pp. 374–384.

[56] Discord Inc. *Discord — Chat for Communities and Friends*. 2019. URL: https://discordapp.com/ (visited on 04/16/2020).

[57] HubPages Inc. *Charades: Topic Ideas, Word Lists, and How to Play | HobbyLark*. 2020. URL: https://hobbylark.com/party-games/Charades-Ideas (visited on 04/16/2020).

[58] Jason Jerald. *The VR book: Human-centered design for virtual reality*. Morgan & Claypool, 2015.

[59] Haiyang Jin, Qing Chen, Zhixian Chen, Ying Hu, and Jianwei Zhang. "Multi-LeapMotion sensor based demonstration for robotic refine tabletop object manipulation task". In: *CAAI Transactions on Intelligence Technology* 1.1 (2016), pp. 104–113.

[60] *Joint Filtering | Microsoft Docs*. URL: https://docs.microsoft.com/en-us/previous-versions/windows/kinect-1.8/jj131024(v=ieb.10) (visited on 05/12/2020).

[61] Sungchul Jung and Charles E. Hughes. "The Effects of Indirect Real Body Cues of Irrelevant Parts on Virtual Body Ownership and Presence". In: *Proceedings of the 26th International Conference on Artificial Reality and Telexistence and the 21st Eurographics Symposium on Virtual Environments*. ICAT-EGVE '16. Little Rock, Arkansas: Eurographics Association, 2016, pp. 107–114. ISBN: 978-3-03868-012-3. DOI: 10.2312/egve.20161442. URL: https://doi.org/10.2312/egve.20161442.

[62] Sungchul Jung, Christian Sandor, Pamela J. Wisniewski, and Charles E. Hughes. "RealME: The Influence of Body and Hand Representations on Body Ownership and Presence". In: *Proceedings of the 5th Symposium on Spatial User Interaction*. SUI '17. Brighton, United Kingdom: ACM, 2017, pp. 3–11. ISBN: 978-1-4503-5486-8. DOI: 10.1145/3131277.3132186. URL: http://doi.acm.org/10.1145/3131277.3132186.

[63] Sungchul Jung, Pamela J Wisniewski, and Charles E. Hughes. "In Limbo: The Effect of Gradual Visual Transition between Real and Virtual on Virtual Body Ownership Illusion and Presence". In: *Proceedings of the 25th IEEE Conference on Virtual Reality and 3D User Interfaces, IEEE VR 2018*. Reutlingen, Germany: IEEE, 2018, pp. 267–272.

[64] Suttipong Kaenchan, Pornchai Mongkolnam, Bunthit Watanapa, and Sasipa Sathienpong. "Automatic multiple Kinect cameras setting for simple walking posture analysis". In: *2013 International Computer Science and Engineering Conference (ICSEC)*. IEEE, Sept. 2013, pp. 245–249. ISBN: 978-1-4673-5324-3. DOI: 10.1109/ICSEC.2013.6694787. URL: http://ieeexplore.ieee.org/document/6694787/.

[65] Konstantina Kilteni, Ilias Bergstrom, and Mel Slater. "Drumming in Immersive Virtual Reality: The Body Shapes the Way We Play". In: *IEEE Transactions on Visualization and Computer Graphics* 19.4 (Apr. 2013), pp. 597–605. ISSN: 1077-2626. DOI: 10.1109/TVCG.2013.29. URL: http://ieeexplore.ieee.org/document/6479188/.

[66] Konstantina Kilteni, Raphaela Groten, and Mel Slater. "The sense of embodiment in virtual reality". In: *Presence: Teleoperators and Virtual Environments* 21.4 (2012), pp. 373–387.

[67] David Kim, Otmar Hilliges, Shahram Izadi, Alex D Butler, Jiawen Chen, Iason Oikonomidis, and Patrick Olivier. "Digits: freehand 3D interactions anywhere using a wrist-worn gloveless sensor". In: *Proceedings of the 25th annual ACM symposium on User interface software and technology*. ACM. 2012, pp. 167–176.

[68] Junghwan KIM, Inwoong LEE, Jongyoo KIM, and Sanghoon LEE. "Implementation of an Omnidirectional Human Motion Capture System Using Multiple Kinect Sensors". In: *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* E98.A.9 (2015), pp. 2004–2008. ISSN: 0916-8508. DOI: 10.1587/transfun.E98.A.2004. URL: https://www.jstage.jst.go.jp/article/transfun/E98.A/9/E98.A%7B%5C_%7D2004/%7B%5C_%7Darticle.

[69] Beom Kwon, Junghwan Kim, Kyoungoh Lee, Yang Koo Lee, Sangjoon Park, and Sanghoon Lee. "Implementation of a Virtual Training Simulator Based on 360° Multi-View Human Action Recognition". In: *IEEE Access* 5 (2017), pp. 12496–12511. ISSN: 2169-3536. DOI: 10.1109/

ACCESS.2017.2723039. URL: http://ieeexplore.ieee.org/document/7968260/.

[70] Joung Huem Kwon, John Powell, and Alan Chalmers. "How level of realism influences anxiety in virtual reality environments for a job interview". In: *International journal of human-computer studies* 71.10 (2013), pp. 978–987.

[71] Marc Erich Latoschik, Jean Luc Lugriny, and Daniel Rothz. "FakeMi: A fake mirror system for avatar embodiment studies". In: *Proceedings of the ACM Symposium on Virtual Reality Software and Technology, VRST* 02-04-Nove (2016), pp. 73–76. DOI: 10.1145/2993369.2993399.

[72] Leap_Motion. *Leap motion_HTC Vive Setup*. URL: https://developer.leapmotion.com/vr-setup/vive.

[73] Oculus Lipsync. *Viseme Reference*. 2019. URL: https://developer.oculus.com/documentation/unity/audio-ovrlipsync-viseme-reference/ (visited on 04/14/2020).

[74] Ultrahaptics Ltd. *Unity, Leap Motion Developer*. 2017. URL: https://developer.leapmotion.com/unity%7B%5C#%7D5436356 (visited on 06/25/2019).

[75] Lara Maister, Mel Slater, Maria V Sanchez-Vives, and Manos Tsakiris. "Changing bodies changes minds: owning another body affects social cognition". In: *Trends in cognitive sciences* 19.1 (2015), pp. 6–12.

[76] MakeHuman. *www.makehumancommunity.org*. 2018. URL: http://www.makehumancommunity.org/ (visited on 06/25/2019).

[77] David Matsumoto, Mark G Frank, and Hyi Sung Hwang. *Nonverbal communication: Science and applications*. Sage Publications, 2012.

[78] Lynn McAtamney and E Nigel Corlett. "RULA: a survey method for the investigation of work-related upper limb disorders". In: *Applied ergonomics* 24.2 (1993), pp. 91–99.

[79] Steven McCornack and Joseph Ortiz. *Choices & Connections: An Introduction to Communication*. Macmillan Higher Education, 2019.

[80] Joshua McVeigh-Schultz, Anya Kolesnichenko, and Katherine Isbister. "Shaping Pro-Social Interaction in VR: An Emerging Design Framework". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–12.

[81] Thomas Metzinger. "Empirical perspectives from the self-model theory of subjectivity: a brief summary with examples". In: *Progress in brain research* 168 (2007), pp. 215–278.

[82] Marvin Minsky. "Telepresence". In: (1980).

[83] Roberto A Montano Murillo, Sriram Subramanian, and Diego Martinez Plasencia. "Erg-O: ergonomic optimization of immersive virtual environments". In: *Proceedings of the 30th annual ACM symposium on user interface software and technology*. 2017, pp. 759–771.

[84] Leonel Morgado, Bernardo Cardoso, Fausto de Carvalho, Luıs Fernandes, Hugo Paredes, Luıs Barbosa, Benjamim Fonseca, Paulo Martins, and Ricardo Rodrigues Nunes. "Separating gesture detection and application control concerns with a multimodal architecture". In: (2015), pp. 1548–1553.

[85] Leap Motion. *Leap Motion API in C*. URL: https://developer.leapmotion.com/documentation/v4/index.html (visited on 04/11/2020).

[86] Björn Müller, Winfried Ilg, Martin A Giese, and Nicolas Ludolph. "Improved Kinect sensor based motion capturing system for gait assessment". In: (2017). DOI: 10.1101/098863. URL: http://biorxiv.org/content/early/2017/01/10/098863.abstract.

[87] Dana S. Nau. *Rules for Charades*. 2000. URL: https://www.cs.umd.edu/users/nau/misc/charades.html (visited on 04/16/2020).

[88] Alexander Nilsson, Ann-Sofie Axelsson, Ilona Heldal, and Ralph Schroeder. "The long-term uses of shared virtual environments: An exploratory study". In: *The social life of avatars*. Springer, 2002, pp. 112–126.

[89] Kristine Nowak. "Defining and differentiating copresence, social presence and presence as transportation". In: *Presence 2001 Conference, Philadelphia, PA*. Citeseer. 2001, pp. 1–23.

[90] Kristine L. Nowak and Frank Biocca. "The Effect of the Agency and Anthropomorphism on Users' Sense of Telepresence, Copresence, and Social Presence in Virtual Environments". In: *Presence: Teleoperators and Virtual Environments* 12.5 (Oct. 2003), pp. 481–494. ISSN: 1054-7460. DOI: 10.1162/105474603322761289. URL: http://www.mitpressjournals.org/doi/10.1162/105474603322761289.

[91] Soo Youn Oh, Jeremy Bailenson, Nicole Krämer, and Benjamin Li. "Let the Avatar Brighten Your Smile: Effects of Enhancing Facial Expressions in Virtual Environments". In: *PLOS ONE* 11.9 (2016), pp. 1–18. DOI: 10.1371/journal.pone.0161794. URL: https://doi.org/10.1371/journal.pone.0161794.

[92] *OpenCV: Pose Estimation*. URL: https://docs.opencv.org/3.1.0/d7/d53/tutorial%7B%5C_%7Dpy%7B%5C_%7Dpose.html (visited on 05/12/2020).

[93] Ye Pan and Anthony Steed. "The impact of self-avatars on trust and collaboration in shared virtual environments". In: *PloS one* 12.12 (2017), e0189078.

[94] PCL. *PCL Homepage*. URL: http://pointclouds.org/ (visited on 05/12/2020).

[95] Giuseppe Placidi, Luigi Cinque, Andrea Petracca, Matteo Polsinelli, and Matteo Spezialetti. "A Virtual Glove System for the Hand Rehabilitation based on Two Orthogonal LEAP Motion Controllers." In: *ICPRAM*. 2017, pp. 184–192.

[96] Holger Regenbrecht, Jonny Collins, and Simon Hoermann. "A leap-supported, hybrid AR interface approach". In: *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration*. 2013, pp. 281–284.

[97] Ronald E Rice. "Media appropriateness: Using social presence theory to compare traditional and new organizational media". In: *Human communication research* 19.4 (1993), pp. 451–484.

[98] RootMotion. *Final IK - RootMotion*. 2019. URL: http://www.root-motion.com/final-ik.html (visited on 06/25/2019).

[99] Daniel Roth, Gary Bente, Peter Kullmann, David Mal, Chris Felix Purps, Kai Vogeley, and Marc Erich Latoschik. "Technologies for Social Augmentations in User-Embodied Virtual Reality". In: *25th ACM Symposium on Virtual Reality Software and Technology*. 2019, pp. 1–12.

[100] Daniel Roth, Carola Bloch, Josephine Schmitt, Lena Frischlich, Marc Erich Latoschik, and Gary Bente. "Perceived Authenticity, Empathy, and Pro-social Intentions evoked through Avatar-mediated Self-disclosures". In: *Proceedings of Mensch und Computer 2019*. 2019, pp. 21–30.

[101] Daniel Roth, Jean Luc Lugrin, Julia Buser, Gary Bente, Arnulph Fuhrmann, and Marc Erich Latoschik. "A simplified inverse kinematic approach for embodied VR applications". In: *Proceedings - IEEE Virtual Reality* 2016-July (2016), pp. 275–276. DOI: 10.1109/VR.2016.7504760.

[102] Daniel Roth, David Mal, Christian Felix Purps, Peter Kullmann, and Marc Erich Latoschik. "Injecting Nonverbal Mimicry with Hybrid Avatar-Agent Technologies: A NaÏVe Approach". In: *Proceedings of the Symposium on Spatial User Interaction*. SUI '18. Berlin, Germany: ACM, 2018, pp. 69–73. ISBN: 978-1-4503-5708-1. DOI: 10.1145/3267782.3267791. URL: http://doi.acm.org/10.1145/3267782.3267791.

[103] Daniel Roth, Kristoffer Waldow, Marc Erich Latoschik, Arnulph Fuhrmann, and Gary Bente. "Socially immersive avatar-based communication". In: *2017 IEEE Virtual Reality (VR)*. IEEE. 2017, pp. 259–260.

[104] Rug.OSC. *Homepage of Rug.OSC*. URL: https://bitbucket.org/rugcode/rug.osc/src/master/ (visited on 04/11/2020).

[105] Eva-Lotta Salinäs. "Collaboration in multi-modal virtual worlds: comparing touch, text, voice and video". In: *The social life of avatars*. Springer, 2002, pp. 172–187.

[106] Mohammed Samir, Ehsan Golkar, and Ashrani Aizzuddin Abd. Rahni. "Comparison between the Kinect™ V1 and Kinect™ V2 for respiratory motion tracking". In: *2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. IEEE, Oct. 2015, pp. 150–155. ISBN: 978-1-4799-8996-6. DOI: 10.1109/ICSIPA.2015.7412180. URL: http://ieeexplore.ieee.org/document/7412180/.

[107] Ralph Schroeder. "Social interaction in virtual environments: Key issues, common themes, and a framework for research". In: *The social life of avatars*. Springer, 2002, pp. 1–18.

[108] Ralph Schroeder. *The social life of avatars: Presence and interaction in shared virtual environments*. Springer Science & Business Media, 2012.

[109] Thomas Schubert, Frank Friedmann, and Holger Regenbrecht. "The experience of presence: Factor analytic insights". In: *Presence: Teleoperators & Virtual Environments* 10.3 (2001), pp. 266–281.

[110] John Short, Ederyn Williams, and Bruce Christie. *The social psychology of telecommunications*. John Wiley & Sons, 1976.

[111] Mel Slater and Maria V Sanchez-Vives. "Enhancing our lives with immersive virtual reality". In: *Frontiers in Robotics and AI* 3 (2016), p. 74.

[112] Mel Slater, Bernhard Spanlang, Maria V Sanchez-Vives, and Olaf Blanke. "First person experience of body transfer in virtual reality". In: *PloS one* 5.5 (2010).

[113] Mel Slater and Anthony Steed. "Meeting People Virtually: Experiments in Shared Virtual Environments". In: *The Social Life of Avatars: Presence and Interaction in Shared Virtual Environments*. Ed. by Ralph Schroeder. London: Springer London, 2002, pp. 146–171. ISBN: 978-1-4471-0277-9.

DOI: 10.1007/978-1-4471-0277-9_9. URL: https://doi.org/10.1007/978-1-4471-0277-9%7B%5C_%7D9.

[114] Mel Slater and Martin Usoh. "Body centred interaction in immersive virtual environments". In: *Artificial life and virtual reality* 1.1994 (1994), pp. 125–148.

[115] Harrison Jesse Smith and Michael Neff. "Communication Behavior in Embodied Virtual Reality". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18* (2018), pp. 1–12. DOI: 10.1145/3173574.3173863. URL: http://dl.acm.org/citation.cfm?doid=3173574.3173863.

[116] Harrison Jesse Smith and Michael Neff. "Communication Behavior in Embodied Virtual Reality". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada: ACM, 2018, 289:1–289:12. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3173863. URL: http://doi.acm.org/10.1145/3173574.3173863.

[117] Bernhard Spanlang, Jean-Marie Normand, David Borland, Konstantina Kilteni, Elias Giannopoulos, AusiÃ s PomÃ©s, Mar GonzÃ¡lez-Franco, Daniel Perez-Marcos, Jorge Arroyo-Palacios, Xavi Navarro Muncunill, and Mel Slater. "How to Build an Embodiment Lab: Achieving Body Representation Illusions in Virtual Reality". In: *Frontiers in Robotics and AI* 1.November (Nov. 2014), pp. 1–22. ISSN: 2296-9144. DOI: 10.3389/frobt.2014.00009. URL: http://journal.frontiersin.org/article/10.3389/frobt.2014.00009/abstract.

[118] Misha Sra and Chris Schmandt. "MetaSpace II: Object and full-body tracking for interaction and navigation in social VR". In: *arXiv preprint arXiv:1512.02922* (2015).

[119] Anthony Steed and Ralph Schroeder. "Collaboration in Immersive and Non-immersive Virtual Environments". In: *Immersed in Media*. Cham: Springer International Publishing, 2015, pp. 263–282. ISBN: 9783319101903. DOI: 10.1007/978-3-319-10190-3_11. URL:

http://link.springer.com/10.1007/978-3-319-10190-3%7B%5C_%7D11.

[120] Anthony Steed, Mel Slater, Amela Sadagic, Adrian Bullock, and Jolanda Tromp. "Leadership and collaboration in shared virtual environments". In: *Proceedings IEEE Virtual Reality (Cat. No. 99CB36316)*. IEEE. 1999, pp. 112–115.

[121] Jonathan Steuer. "Defining Virtual Reality: Dimensions Determining Telepresence". In: *Journal of Communication* 42.4 (1992), pp. 73–93. ISSN: 14602466. DOI: 10.1111/j.1460-2466.1992.tb00812.x.

[122] Crazy Minnow Studio. *SALSA LipSync V1*. 2014. URL: https://assetstore.unity.com/packages/tools/animation/salsa-lipsync-suite-148442 (visited on 06/25/2019).

[123] Crazy Minnow Studio. *SALSA LipSync V2 | Animation | Unity Asset Store*. 2019. URL: https://assetstore.unity.com/packages/tools/animation/salsa-lipsync-suite-148442%7B%5C#%7Dcontent (visited on 04/14/2020).

[124] K Swinth and Jim Blascovich. "Perceiving and responding to others: Human-human and human-computer social interaction in collaborative virtual environments". In: *Proceedings of the 5th Annual International Workshop on PRESENCE*. Vol. 392. 2002.

[125] Unity Technologies. *Unity Real-Time Development Platform | 3D, 2D VR & AR Visualizations*. 2017. URL: https://unity.com/ (visited on 05/15/2020).

[126] "The effect of avatar realism in immersive social virtual realities". In: (2017), pp. 1–10. DOI: 10.1145/3139131.3139156.

[127] "The Effect of Behavioral Realism and Form Realism of Real-Time Avatar Faces on Verbal Disclosure, Nonverbal Disclosure, Emotion Recognition, and Copresence in Dyadic Interaction". In: *Presence: Teleoperators and Virtual Environments* 15.4 (Aug. 2006), pp. 359–372. ISSN:

1054-7460. DOI: 10.1162/pres.15.4.359. URL: http://www.mitpressjournals.org/doi/10.1162/pres.15.4.359.

[128] "Toward Safe Human Robot Collaboration by Using Multiple Kinects Based Real-time Human Tracking". In: *Journal of Computing and Information Science in Engineering* 14.1 (2014), p. 011006. ISSN: 1530-9827. DOI: 10.1115/1.4025810.

[129] Manos Tsakiris. "My body in the brain: a neurocognitive model of body-ownership". In: *Neuropsychologia* 48.3 (2010), pp. 703–712.

[130] *UniOSC | UniOSC – the OSC solution for Unity3d*. URL: http://uniosc.monoflow.org/ (visited on 05/12/2020).

[131] Geoffrey C. Urbaniak and Scott Plous. *Research Randomizer*. 1997. URL: https://www.randomizer.org/ (visited on 06/25/2019).

[132] Daniela Villani, Claudia Repetto, Pietro Cipresso, and Giuseppe Riva. "May I experience more presence in doing the same thing in virtual reality than in reality? An answer from a simulated job interview". In: *Interacting with Computers* 24.4 (2012), pp. 265–272.

[133] Daniela Villani, Chiara Rotasperti, Pietro Cipresso, Stefano Triberti, Claudia Carissoli, and Giuseppe Riva. "Assessing the emotional state of job applicants through a virtual reality simulation: a psycho-physiological study". In: *eHealth 360°*. Cham: Springer, 2017, pp. 119–126.

[134] Joshua Wainer, David J Feil-Seifer, Dylan A Shell, and Maja J Mataric. "The role of physical embodiment in human-robot interaction". In: *RO-MAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE. 2006, pp. 117–122.

[135] Joseph B Walther. "Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction". In: *Communication research* 23.1 (1996), pp. 3–43.

[136] Steue Whittaker. "Theories and Methods in Mediated Communication: Steve Whittaker". In: *Handbook of discourse processes*. London, UK: Routledge, 2003, pp. 246–289.

[137] Bob G Witmer and Michael J Singer. "Measuring presence in virtual environments: A presence questionnaire". In: *Presence* 7.3 (1998), pp. 225–240.

[138] Yuanjie Wu, Simon Hoermann, and Robert W Lindeman. "Towards Robust 3D Skeleton Tracking Using Data Fusion from Multiple Depth Sensors". In: *2018 10th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)* (2018), pp. 3–6.

[139] Yuanjie Wu, Yu Wang, Sungchul Jung, Simon Hoermann, and Robert W. Lindeman. "Exploring the Use of a Robust Depth-Sensor-Based Avatar Control System and Its Effects on Communication Behaviors". In: *25th ACM Symposium on Virtual Reality Software and Technology*. VRST '19. Parramatta, NSW, Australia: Association for Computing Machinery, 2019. ISBN: 9781450370011. DOI: 10.1145/3359996.3364267. URL: https://doi.org/10.1145/3359996.3364267.

[140] Yuanjie Wu, Yu Wang, Sungchul Jung, Simon Hoermann, and Robert W. Lindeman. "Towards an articulated avatar in VR: Improving body and hand tracking using only depth cameras". In: *Entertainment Computing* 31 (2019), p. 100303. ISSN: 1875-9521. DOI: https://doi.org/10.1016/j.entcom.2019.100303. URL: http://www.sciencedirect.com/science/article/pii/S1875952119300138.

[141] Nick Yee and Jeremy Bailenson. "The Proteus effect: The effect of transformed self-representation on behavior". In: *Human communication research* 33.3 (2007), pp. 271–290.

[142] Zhengyou Zhang. "A flexible new technique for camera calibration". In: *IEEE Transactions on pattern analysis and machine intelligence* 22.11 (2000), pp. 1330–1334.

[143] Zhengyou Zhang. "Iterative point matching for registration of free-form curves and surfaces". In: *International journal of computer vision* 13.2 (1994), pp. 119–152. DOI: 10.1007/BF01427149. URL: https://doi.org/10.1007/BF01427149.

[144] Zhengyou Zhang. "Microsoft kinect sensor and its effect". In: *IEEE Multimedia* 19.2 (2012), pp. 4–10. ISSN: 1070986X. DOI: 10.1109/MMUL.2012.24.

Appendix A

Experiment documents

This section presents the ethic approvals and questionnaires used in the user studies in Chapter 5 and Chapter 7.

All the data and documents of user studies during my PhD research are here: https://drive.google.com/drive/folders/19DALfzPL4KA5h8Aor6EkFYE1v0G30Iph?usp=sharing

HUMAN ETHICS COMMITTEE

Secretary, Rebecca Robinson
Telephone: +64 03 369 4588, Extn 94588
Email: human-ethics@canterbury.ac.nz

Ref: HEC 2019/16/LR-PS

27 May 2019

Yuanjie Wu
HIT Lab NZ
UNIVERSITY OF CANTERBURY

Dear Yuanjie

Thank you for submitting your low risk application to the Human Ethics Committee for the research proposal titled "Exploring the Effect of Highly Expressive Avatars on Player Behaviour in a Virtual Interview".

I am pleased to advise that this application has been reviewed and approved.

Please note that this approval is subject to the incorporation of the amendments you have provided in your email of 17th May 2019.

With best wishes for your project.

Yours sincerely

Dr Dean Sutherland
Chair, Human Ethics Committee

HUMAN ETHICS COMMITTEE

Secretary, Rebecca Robinson
Telephone: +64 03 369 4588, Extn 94588
Email: human-ethics@canterbury.ac.nz

Ref: HEC 2019/47/LR-PS

2 December 2019

Yuanjie Wu
HIT Lab NZ
UNIVERSITY OF CANTERBURY

Dear Yuanjie

Thank you for submitting your low risk application to the Human Ethics Committee for the research proposal titled "Exploring the Effect of Expressive Avatars on Social Behaviour in Collaborative Virtual Environments".

I am pleased to advise that this application has been reviewed and approved.

Please note that this approval is subject to the incorporation of the amendments you have provided in your email of 26th November 2019.

With best wishes for your project.

Yours sincerely

Dr Dean Sutherland
Chair, Human Ethics Committee

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Age

Gender

- Male
- Female
- Other
- Choose not to answer

Have you used a VR headset before?

- Never
- A few times a year
- A few times a month
- A few times a week
- Daily

Have you ever tried Facebook Space, High Fidelity or VR chat or any other social VR?

- Never
- A few times a year
- A few times a month
- A few times a week
- Daily

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Please indicate, whether or not each statement applies to your experience. You can use the whole range of answers. There are no right or wrong answers; only your opinion counts.

You will notice that some questions are very similar to each other. This is necessary for statistical reasons.

Please answer all question with reference to the VR session you just completed.

In the computer generated world, I had a sense of "being there".

not at all ○ ○ ○ ○ ○ ○ ○ very much

Somehow I felt that the virtual world surrounded me.

fully disagree ○ ○ ○ ○ ○ ○ ○ fully agree

I felt like I was just perceiving pictures.

fully disagree ○ ○ ○ ○ ○ ○ ○ fully agree

I did not feel present in the virtual space.

did not feel present ○ ○ ○ ○ ○ ○ ○ felt present

I had a sense of acting in the virtual space, rather than operating something from outside.

fully disagree ○ ○ ○ ○ ○ ○ ○ fully agree

I felt present in the virtual space.

fully disagree ○ ○ ○ ○ ○ ○ ○ fully agree

How aware were you of the real world surrounding while navigating in the virtual world? (i.e., sounds, room temperature, other people, etc.)?

extremely aware ○ ○ ○ ○ ○ ○ ○ not aware at all

I was not aware of my real environment.

fully disagree ○ ○ ○ ○ ○ ○ ○ fully agree

I still paid attention to the real environment.

fully disagree ○ ○ ○ ○ ○ ○ ○ fully agree

I was completely captivated by the virtual world.

fully disagree ○ ○ ○ ○ ○ ○ ○ fully agree

How real did the virtual world seem to you?

completely real ○ ○ ○ ○ ○ ○ ○ not real at all

How much did your experience in the virtual environment seem consistent with your real world experience ?

not consistent ○ ○ ○ ○ ○ ○ ○ very consistent

How real did the virtual world seem to you?

about as real as an imagined world ○ ○ ○ ○ ○ ○ ○ indistinguishable from the real world

The virtual world seemed more realistic than the real world.

fully disagree ○ ○ ○ ○ ○ ○ ○ fully agree

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Please select your level of agreement with the following statements:
"During the experiment there were moments in which…

I felt as if the virtual body I saw when I looked down was my body"

Strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

It felt as if the virtual body I saw was someone else"

Strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

It seemed as if I might have more than one body"

Strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

It felt like I could control the virtual body as if it was my own body"

Strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

The movements of the virtual body were caused by my movements"

Strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

I felt as if the movements of the virtual body were influencing my own movements"

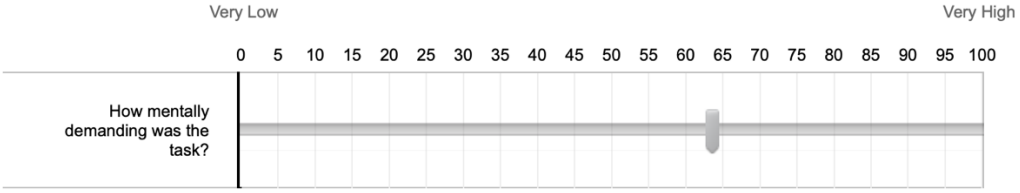
Strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

I felt as if the virtual body was moving by itself"

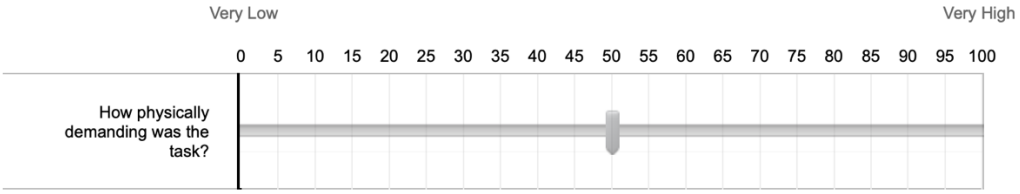
Strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

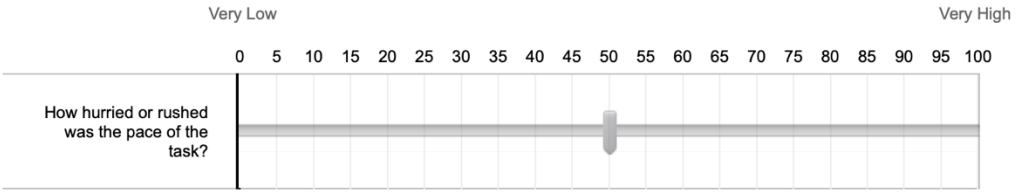
Mental Demand



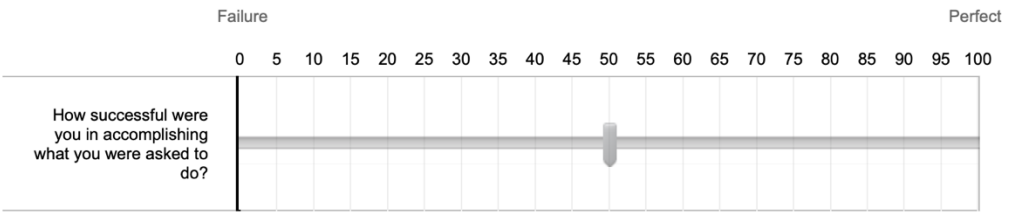
Physical Demand



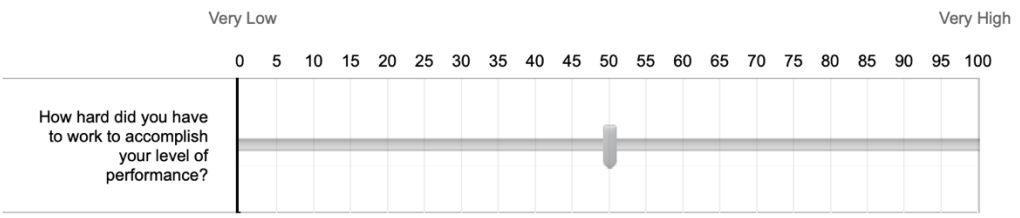
Temporal Demand



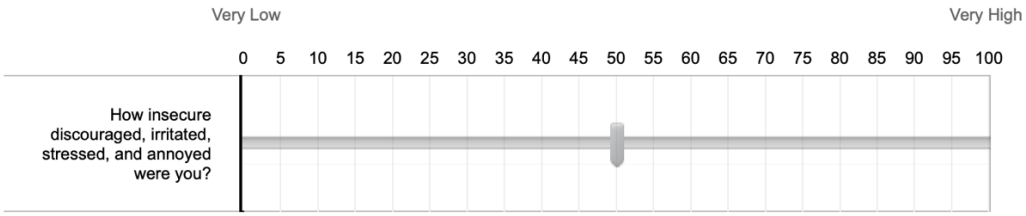
Performance



Effort



Frustration



UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

I think that I would like to use this system frequently.

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

I found the system unnecessarily complex.

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

I thought the system was easy to use.

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

I think that I would need the support of a technical person to be able to use this system.

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

I found the various functions in this system were well integrated.

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

I thought there was too much inconsistency in this system.

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

I would imagine that most people would learn to use this system very quickly.

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

I found the system very cumbersome to use.

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

I felt very confident using the system.

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

I needed to learn a lot of things before I could get going with this system.

Strongly disagree ○ ○ ○ ○ ○ Strongly agree

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Think about what you saw when you watched the replay of your interview.
How realistic was...

0 10 20 30 40 50 60 70 80 90 100

... your non-verbal behavior: body posture and hand gestures?

... your verbal behavior?

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

1. Which VR system was easier to use?

- ○ Multi-sensor integrated tracking system
- ○ Semi-automated tracking system with IK

2. Which VR system do you prefer?

- ○ Multi-sensor integrated tracking system
- ○ Semi-automated tracking system with IK

Any comments on your experience?

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Participant number:

Age

Gender

- Male
- Female
- Other
- Choose not to answer

Do you have a corrected vision?

- No, I have good eyesight
- Yes, I wear contact lenses or glasses

What is your English speaking level?

- I am not a native English speaker and I can not speak English fluently
- I am not a native English speaker but I can speak English fluently
- I am a native English speaker

What is your relationship with another player?

- We do not know each other
- We know each other but we are only classmates or colleagues
- We know each other and we are friends

Have you used a VR headset before?

- Never
- A few times a year
- A few times a month
- A few times a week
- Daily

Have you ever tried Facebook Space, High Fidelity or VR chat or any other social VR?

- ◯ Never
- ◯ A few times a year
- ◯ A few times a month
- ◯ A few times a week
- ◯ Daily

Have you ever played the **charades game** before? If yes, what is your level of expertise?

- ◯ Never
- ◯ I am a beginner
- ◯ I am an intermediate player
- ◯ I am an expert

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Please indicate, whether or not each statement applies to your experience. You can use the whole range of answers. There are no right or wrong answers; only your opinion counts.

You will notice that some questions are very similar to each other. This is necessary for statistical reasons.

Please answer all question with reference to the VR session you just completed.

I did not want a deeper relationship with my interaction partner.

strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

I wanted to maintain a sense of distance between us.

strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

I was unwilling to share personal information with my interaction partner.

strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

I wanted to make the conversation more intimate.

strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

I tried to create a sense of closeness between us.

strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

I was interested in talking to my interaction partner.

strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

My interaction partner was intensely involved in our interaction.

strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

My interaction partner seemed to find our interaction stimulating.

strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

My interaction partner communicated coldness rather than warmth.

strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

My interaction partner created a sense of distance between us.

strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

My interaction partner seemed detached during our interaction.

strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

My interaction partner was unwilling to share personal information with me.

strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

My interaction partner made our conversation seem intimate.

strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

My interaction partner created a sense of distance between us.

strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

My interaction partner created a sense of closeness between us.

strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

My interaction partner acted bored by our conversation.

strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

My interaction partner was interested in talking to me.

strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

My interaction partner showed enthusiasm while talking to me.

strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Please select your level of agreement with the following statements:

To what extent did you feel able to assess your partner's reactions to what you said?

Not able to assess reactions — Able to assess reactions

0 10 20 30 40 50 60 70 80 90 100

To what extent was this like a face-to-face meeting?

Not like face to face at all — A lot like face to face

0 10 20 30 40 50 60 70 80 90 100

To what extent was this like you were in the same room with your partner?

Not like being in the same room at all — A lot like being in the same room

0 10 20 30 40 50 60 70 80 90 100

To what extent did your partner seem "real"?

Not real at all — Very real

0 10 20 30 40 50 60 70 80 90 100

How likely is it that you would choose to use this system of interaction for a meeting in which you wanted to persuade others of something?

Not likely at all — Very likely

0 10 20 30 40 50 60 70 80 90 100

To what extent did you feel you could get to know someone that you met only through this system?

Not at all Very well

0 10 20 30 40 50 60 70 80 90 100

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

How strongly do you agree or disagree with the following statements about your partner?

I like my partner.

strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

I would get along well with my partner.

strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

I would enjoy a casual conversation with my partner.

strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

My partner is the type of person I could become close friends with.

strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

My partner is a good listener.

strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

My partner is friendly.

strongly disagree	disagree	somewhat disagree	neither agree nor disagree	somewhat agree	agree	strongly agree
○	○	○	○	○	○	○

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

1. Which VR system was most helpful when you were describing words to your partner?

- ○ Low expressiveness avatar system (Controller based system)
- ○ High expressiveness avatar system (Sensors based system)

2. Which VR system do you prefer?

- ○ Low expressiveness avatar system (Controller based system)
- ○ High expressiveness avatar system (Sensors based system)

Any comments on your experience?